Privacy-preserving Sparse Generalized Eigenvalue Problem

Lijie Hu KAUST
Zihang Xiang KAUST
Jiabin Liu Beijing Institute of Technology
Di Wang KAUST

Abstract

In this paper we study the (sparse) Generalized Eigenvalue Problem (GEP), which arises in a number of modern statistical learning models, such as principal component analysis (PCA), canonical correlation analysis (CCA), Fisher's discriminant analysis (FDA) and sliced inverse regression (SIR). We provide the first study on GEP in the differential privacy (DP) model under both deterministic and stochastic settings. In the low dimensional case, we provide a ρ-Concentrated DP (CDP) method namely DP-Rayleigh Flow and show if the initial vector is close enough to the optimal vector, its output has an ℓ_2-norm estimation error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ (under some mild assumptions), where d is the dimension and n is the sample size. Next, we discuss how to find such an initial parameter privately. In the high dimensional sparse case where $d \gg n$, we propose the DP-Truncated Rayleigh Flow method whose output could achieve an error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ for various statistical models, where s is the sparsity of the underlying parameter. Moreover, we show that these errors in the stochastic setting are optimal up to a factor of $\text{Poly}(\log n)$ by providing the lower bounds of PCA and SIR under the statistical setting and in the CDP model. Finally, to give a separation between ϵ-DP and ρ-CDP for GEP, we also provide the lower bound $\Omega(\frac{d}{n} + \frac{d^2}{n^2\epsilon^2})$ and $\Omega(\frac{s\log d}{n} + \frac{s^2\log^2 d}{n^2\epsilon^2})$ of private minimax risk for PCA, under the statistical setting and ϵ-DP model, in low and high dimensional sparse case respectively. [1]

[1]The first two authors contributed equally.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

(Sparse) generalized eigenvalue problem (GEP) has received much attention recently as it arises in a number of standard and modern statistical learning models, including (sparse) principal component analysis (PCA), (sparse) Fisher's discriminant analysis (FDA), and (sparse) canonical correlation analysis (CCA), which have enormous applications in biomedicine [Liu and Altman, 2015], biomedical imaging [Strickert et al., 2009] and genomics [Parkhomenko et al., 2009].

The wide applications of GEP also present some new challenges to this problem. Particularly, due to the existence of sensitive data (such as biomedical images) and their distributed nature in many applications like biomedicine and genomics, it is often challenging to preserve the privacy of such data as they are extremely difficult to aggregate and learn from. One promising direction is to use some differentially private mechanisms to conduct the aggregation and learning tasks. Differential Privacy (DP) [Dwork et al., 2006] is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers. To design DP algorithms, previous work always focus on specific statistical models, such as (sparse) PCA, CCA. However, there is no general framework which can solve them all together. As the above problems all can be formulated as a GEP problem, a DP algorithm for (sparse) GEP could simultaneously solve PCA, CCA, FDA etc. However, to our best knowledge, there is no work on the designing DP algorithms for (sparse) GEP and the theoretical behaviors of GEP in DP model is still unknown.

To address the above issues, in this paper, we provide a first study of GEP under the DP constraint, *i.e., DP-GEP*, under both low dimension and high dimensional sparse settings. Specifically, our contributions can be summarized as following.

- We first consider DP-GEP in the low dimensional case. Specifically, we propose a ρ-Concentrated DP (CDP) method, namely DP-Rayleigh Flow, and show that if the initial vector is close enough the optimal one, then the output of algorithm could achieve an

ℓ_2-norm estimation error of $\tilde{O}(\frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n}+\frac{d}{n^2\rho})$ in the deterministic and statistical setting respectively (under some mild assumptions), where n is the sample size and d is the dimension of the space. Moreover, we also show that if n is sufficiently large, then we can efficiently find such an initial parameter with ρ-CDP guarantee by reformulating the original GEP problem as a convex programming problem with a LASSO penalty.

- We then consider the problem in the high dimensional case with $d \gg n$, where we assume the underlying parameter is s-sparse. Particularly, we present a method namely DP-Truncated Rayleigh Flow which could achieve an error of $\tilde{O}(\frac{s\log d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n}+\frac{s\log d}{n^2\rho})$ in deterministic and statistical setting respectively, with some initial parameter. As corollaries, we also provide the first theoretical result for CCA, FDA and Sliced Inverse Regression (SIR) in the CDP model.

- We also study the lower bounds of DP-GEP under various settings. We first show that the previous upper bounds in the stochastic setting are optimal up to a factor of $\text{Poly}(\log n)$ by showing the optimal rates of private minimax risk for PCA and SIR in the CDP model. Then we study the ϵ-DP model and show that the private minimax risk for ϵ-DP-PCA is lower bounded by $\Omega(\frac{d}{n}+\frac{d^2}{n^2\epsilon^2})$ and $\Omega(\frac{s\log d}{n}+\frac{s^2\log^2 d}{n^2\epsilon^2})$ in low dimensional and high dimensional setting respectively. Compared with our upper bounds, we can see a separation of the problem in ϵ-DP and CDP. To the best of our knowledge, these are first lower bounds of DP sparse PCA and DP-SIR under statistical setting, which may could used to other problems. Finally, extensive experiments on both synthetic and real-world data support our previous theoretical analysis.

Due to space limit, the full version of some theorems, all proofs and experiments are included in Appendix.

2 RELATED WORK

As we mentioned earlier, there is no previous result on DP-GEP, and there is even no result on DP-FDA and DP-SIR. For DP-CCA, [Imtiaz and Sarwate, 2017] first studies the problem, which is later extended by [Imtiaz and Sarwate, 2019, Shen, 2020] to other settings. However, their algorithms cannot be extended to the high dimensional sparse case and there is no theoretical guarantees for their methods. Below we will focus on the previous results on DP-PCA.

There is a vast number of papers studying PCA under differential privacy, starting from the SULQ framework [Blum et al., 2005, Dwork et al., 2014, Chaudhuri et al., 2013, Gonem and Gilad-Bachrach, 2018, Ge et al., 2018, Balcan et al., 2016]. For DP-PCA in (ϵ,δ)-DP model, [Hardt and Roth, 2013, Balcan et al., 2016, Hardt and Price, 2014] study noisy versions of the power method. [Dwork et al., 2014] considers the deterministic setting and provides the optimal rate of the problem for general K-PCA. However, all these methods cannot be extended to the high dimensional sparse case. For high dimensional sparse PCA, [Ge et al., 2018] studies the problem in the distributed setting and proposed a noisy iterative hard thresholding power method, and [Wang and Xu, 2020] focuses on the problem in the local DP model by showing its upper bound and lower bound. However, these methods are only for PCA and cannot be extended to GEP where here we have an additional constraint which also depends on the dataset. Moreover, the proof of lower bound is also different since it only focuses on the local DP model while in this paper we study the central one.

There are also several papers provide the lower bounds of PCA in central ϵ-DP model. However, all of them are different with ours. Specifically, [Chaudhuri et al., 2012] studies the deterministic setting and shows the lower bound of $\Omega(\frac{d^2}{n^2\epsilon^2})$ for the estimation error, which is later extended by [Kapralov and Talwar, 2013] to general K-PCA case. Compared with their results, we consider the stochastic setting instead and show the lower bound of $\Omega(\frac{d}{n}+\frac{d^2}{n^2\epsilon^2})$. Due to different settings, their proof techniques cannot be used to ours and we use a different technique of proof. [Liu et al., 2022] recently also studies the lower bound of ϵ-DP-PCA under statistical setting. However, their assumptions on the underlying distribution of data are totally different with ours, which indicates that our results are incomparable with theirs. Moreover, they only consider the low dimensional case while we consider both low dimension and high dimensional sparse cases. For PCA in the central (ϵ,δ)-DP model, [Dwork et al., 2014] provides a lower bound of $\Omega(\frac{d\log\frac{1}{\delta}}{n^2\epsilon^2})$ for the problem in the deterministic setting by using the fingerprinting codes while in this paper we provide the lower bound of $\Omega(\frac{d}{n}+\frac{d}{n^2\rho})$ under the stochastic setting and in the CDP model. We also consider the high dimensional sparse case.

3 PRELIMINARIES

Notations: We denote $\lambda_i(Z), \lambda_{\max}(Z), \lambda_{\min}(Z)$ as the i-th, maximal and minimal eigenvalue of matrix Z respectively. And denote the condition number of a positive definite matrix $Z \in \mathbb{R}^{d\times d}$ as $\kappa(Z) = \frac{\lambda_{\max}(Z)}{\lambda_{\min}(Z)}$. Moreover, let λ_j and $\hat{\lambda}_j$ be the j-th generalized eigenvalue of the matrix pairs (A,B) and $(\hat{A},\hat{B})$ respectively. Given an index set $F \subseteq [d]$, let $Z_F \in \mathbb{R}^{|F|\times|F|}$ be the submatrix of Z where the rows and columns are restricted to the set F. We also denote $\rho(Z,s) = \sup_{\|u\|_2=1,\|u\|_0\le s}|u^TZu|$ and $\rho(Z) = \|Z\|_2 = \rho(Z,d)$. For a pair of symmetric ma-

trix (A, B) we denote its Crawford number as $\text{cr}(A, B) = \min_{v:||v||_2=1} \sqrt{(v^T Av)^2 + (v^T Bv)^2} \geq 0$.

In this section, we recall some definitions related to Differnetial Privacy and Generalized Eigenvalue Problem.

Definition 1 (Differential Privacy [Dwork et al., 2006])**.** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data sample, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, we have $\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon}\Pr(\mathcal{A}(D') \in S) + \delta$. When $\delta = 0$ we call the algorithm is ϵ-DP.

Definition 2 (Concentrated DP [Bun and Steinke, 2016])**.** A randomized algorithm $\mathcal{A}$ is ρ-Concentrated DP (CDP) if for all neighboring datasets D, D' and for all $\alpha > 1$ we have $D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D')) \leq \alpha\rho$, where $D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D'))$ is the α-Rényi divergence between $\mathcal{A}(D)$ and $\mathcal{A}(D')$.

Actually, CDP lives between ϵ-DP and (ϵ, δ)-DP:

Lemma 1 ([Bun and Steinke, 2016])**.** For every $\epsilon > 0$, if algorithm $\mathcal{A}$ is ϵ-DP then it will be $\frac{\epsilon^2}{2}$-CDP. If $\mathcal{A}$ is ρ-CDP, then it will be (ϵ, δ)-DP with $\epsilon = \rho + 2\sqrt{\rho \log \frac{1}{\delta}}$.

By the previous lemma, we can see to achieve a given (ϵ, δ)-DP guarantee, it is sufficient to show the algorithm is $\rho = (\sqrt{\epsilon + \log 1/\delta} - \sqrt{\log 1/\delta})^2 \approx \frac{\epsilon^2}{4 \log 1/\delta}$-CDP. **Thus, all ρ-CDP algorithms with their utility in this paper can be transformed to the (ϵ, δ)-DP version with $\log 1/\delta \gg \epsilon$ by simply replacing ρ by $\frac{\epsilon^2}{4 \log 1/\delta}$.**

In this paper, we will mainly use the Gaussian mechanism and the Composition Theorem to guarantee CDP.

Definition 3 (Gaussian Mechanism)**.** Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{\Delta_2^2}{2\rho}\mathbb{I}_d)$, where where $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, *i.e.,* $\Delta_2(q) = \sup_{D \sim d'} ||q(D) - q(D')||_2$. Gaussian mechanism preserves ρ-CDP.

Lemma 2 (Composition Theorem)**.** If $\mathcal{A}$ is an adaptive composition of CDP algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_T$ where $\mathcal{A}_i$ is ρ_i-CDP. Then $\mathcal{A}$ will be ρ-CDP with $\rho = \sum_{i=1}^T \rho_i$.

Definition 4 (GEP [Golub and Van Loan, 1996])**.** The generalized eigenvalues of the symmetric-definite pair (A, B) are denote by $\lambda(A, B) = \{\lambda | \det(A - \lambda B) = 0\}$. If $\lambda \in \lambda(A, B)$ and v is a non-zero vector satisfies $Av = \lambda Bv$, then v is a generalized eigenvector.

Given an n-size data set $D = \{x_1, \cdots, x_n\}$, matrices $\hat{A}$ and $\hat{B}$. The (largest) generalized eigenvalue problem (GEP) of $(\hat{A}, \hat{B})$ is characterized as

$$\tilde{v} = \arg\max_{v \in \mathbb{R}^d} v^T \hat{A} v, \text{ s.t. } v^T \hat{B} v = 1, \tag{1}$$

where $\hat{A} = \hat{A}(D) \in \mathbb{R}^{d \times d}$ and $\hat{B} = \hat{B}(D) \in \mathbb{R}^{d \times d}$ are matrices that (may) dependent on the dataset D. Besides the **deterministic setting**, for some statistical models we also want to study the **stochastic setting** where we assume each record is sampled from some underlying unknown distribution $\mathcal{P}$. And our goal is to solve the following problem based on the data D, where $A = \mathbb{E}[\hat{A}]$ and $B = \mathbb{E}[\hat{B}]$.

$$v^* = \arg\max_{v \in \mathbb{R}^d} v^T A v, \text{ s.t. } v^T B v = 1. \tag{2}$$

In the high dimensional setting, we assume $d \gg n$ and the underlying parameter v^* in (2) or $\tilde{v}$ in (1) has an additional sparsity structure, *i.e.,* we assume $\|v^*\|_0 \leq s$ or $\|\tilde{v}\|_0 \leq s$ for some $s \ll d$. Now the sparse GEP becomes to

$$\tilde{v}_s = \arg\max_{v \in \mathbb{R}^d} v^T \hat{A} v, \text{ s.t. } v^T \hat{B} v = 1, \|v\|_0 \leq s. \tag{3}$$

$$v_s^* = \arg\max_{v \in \mathbb{R}^d} v^T A v, \text{ s.t. } v^T B v = 1, \|v\|_0 \leq s. \tag{4}$$

In the following, we will provide some statistical models that are special cases of (sparse) GEP.

Principal Component Analysis (PCA): Given dataset $D = \{x_1, \cdots, x_2\}$ with each $x_i \in \mathbb{R}^d$, (sparse) PCA can be formulated as (sparse) GEP with $\hat{B} = I_d$ and $\hat{A} = \hat{\Sigma}$ where $\hat{\Sigma}$ is the covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$ with $\mu = \frac{\sum_{i=1}^n x_i}{n}$. In the stochastic setting A is the polulation version of $\hat{A}$, *i.e.,* $A = \mathbb{E}[(x-\mu)(x-\mu)^T]$ with $\mu = \mathbb{E}[x]$.

Canonical Component Analysis (CCA): Given dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ with each $x_i \in \mathbb{R}^{d_1}$ and $y_i \in \mathbb{R}^{d_2}$, (sparse) CCA can be formulated as (sparse) GEP with

$$\hat{A} = \begin{pmatrix} 0 & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{xy} & 0 \end{pmatrix}, \hat{B} = \begin{pmatrix} \hat{\Sigma}_x & 0 \\ 0 & \hat{\Sigma}_y \end{pmatrix},$$

where $\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)(x_i - \mu_x)^T$, $\hat{\Sigma}_y = \frac{1}{n}\sum_{i=1}^n (y_i - \mu_y)(y_i - \mu_y)^T$ and $\hat{\Sigma}_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)^T$ with $\mu_x = \frac{\sum_{i=1}^n x_i}{n}$ and $\mu_y = \frac{\sum_{i=1}^n y_i}{n}$. In the stochastic setting, A and B are the population version of $\hat{A}$ and $\hat{B}$ respectively.

Fisher's Discriminant Analysis (FDA): Given n samples with K different classes, Fisher's discriminant analysis seeks a low dimensional projection of the samples such that the between-class variance is large relative to the with-class variance. Specifically, it could be formulated as GEP with

$$\hat{A} = \frac{1}{n}\sum_{k=1}^K \sum_{i \in C_k} (x_i - \hat{u}_k)(x_i - \hat{u}_k)^T,$$
$$\hat{B} = \frac{1}{n}\sum_{k=1}^K n_k \hat{u}_k \hat{u}_k^T, \hat{u}_k = \sum_{i \in C_k} \frac{x_i}{n_k}, \tag{5}$$

where C_k is the index set for the k-th class, *i.e.,* if $i \in C_k$ then x_i is in the k-th class, and $n_k = |C_k|$.

Sliced Inverse Regression (SIR): In SIR we have the statistical model $Y = f(v_1^T X, \cdots, v_k^T X, \zeta)$, where ζ is some random noise and is independent on X, $f(\cdot)$ is some unknown link function. It has been shown that under some mild assumptions, the space that is spanned by $v_1, \cdots, v_k$ can be identified [Li, 1991]. Particularly, the first leading eigenvector of the subspace that is spanned by $v_1, \cdots, v_k$ can be identified by solving the GEP with A be the covariance matrix of the conditional expectation $\mathbb{E}(X|Y)$ and B as the covariance matrix of X. That is:

$$\begin{aligned}
&\hat{A} = \hat{\Sigma}_{E(X|Y)}, \hat{B} = \hat{\Sigma}_x, \hat{\Sigma}_{E(X|Y)} = \hat{\Sigma}_x - E[\hat{\Sigma}_{(x|y)}] \\
&\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^T, \\
&\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, u_k = \sum_{i \in C_k} \frac{x_i}{n_k} \\
&E[\hat{\Sigma}_{(x|y)}] = \frac{1}{n}\sum_{k=1}^{K}\sum_{x \in C_k} n_k \frac{(x_i - \mu_k)(x_i - \mu_k)^T}{n_k}
\end{aligned} \tag{6}$$

In the following we present the definition of DP-GEP.

Definition 5 (DP-GEP)**.** Given a dataset $D = \{x_1, \cdots, x_n\}$ and its corresponding (sparse) GEP, the goal of Differentially Private GEP (GEP) is finding a private estimator v_{priv} based on some DP algorithm. Moreover, we want our private estimator close enough to the optimal parameter. Specifically, in this paper, we will mainly use the similarity $1 - \frac{\langle v_{\text{priv}}, v^* \rangle}{\|v_{\text{priv}}\|_2 \|v^*\|_2}$ to the measure the closeness. Based on different settings, v^* could be the optimal vector of the problem (1), (2), (3) or (4).

If we denote $\mathbf{E_A} = \mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{E_B} = \mathbf{B} - \hat{\mathbf{B}}$. Then we can see that the deterministic setting is a special case of the stochastic setting with $E_A = E_B = 0$. Thus, in this paper we mainly focus on the stochastic setting. Next we propose several assumptions that will be used throughout the paper. Assumption 1 requires that the Frobenius norm sensitivity of $\hat{A}$ and $\hat{B}$ are bounded by $O(\frac{1}{n})$.

Assumption 1. Given any neighboring datasets D and D'. For $\hat{A}(D) \in \mathbb{R}^{d \times d}$ and $\hat{B}(D) \in \mathbb{R}^{d \times d}$ in problem (1) we assume $\|\hat{A}(D) - \hat{A}(D')\|_F \leq \frac{C_1}{n}$ and $\|\hat{B}(D) - \hat{B}(D')\|_F \leq \frac{C_2}{n}$ for some constants $C_1, C_2 \geq 0$.

The following assumption is to control the norm of the error matrix E_A and E_B in the statistical setting.

Assumption 2. We assume that for any $0 < s \leq d$ we have $\rho(E_A, s), \rho(E_B, s) = O(\sqrt{\frac{s \log d}{n}})$.

It is notable that Assumption 2 is only used in the utility analysis throughout the paper and is only for simplicity, i.e., the privacy guarantees will still hold even Assumption 2 does not hold. Moreover, all of our utility analysis could be extended to general $\rho(E_A), \rho(E_B), \rho(E_A, s), \rho(E_B, s)$, see Appendix for details. In the following we show that the above assumptions hold (with high probability) for all the statistical models we mentioned previously if each data sample $\|x_i\|_2 \leq 1$. Thus, we can see these two assumptions are mild.

Theorem 1. If each $\|x_i\|_2 \leq 1$ for $i \in [n]$, then PCA, CCA, FDA and SIR all satisfy Assumption 1. Moreover, PCA, CCA and SIR satisfy Assumption 2 (with high probability).

4 LOW DIMENSION CASE

In this section we consider the low dimension case, *i.e.,* problem (1) and (2). To illustrate our idea, we first review the classical method for GEP by using Rayleigh's quotient [Parlett, 1998]. Specifically, problem (1) can be rewritten as

$$\max_{v \in \mathbb{R}^d} J(v) = \frac{v^T \hat{A} v}{v^T \hat{B} v}, \tag{7}$$

where the objective function could be seen as the generalized Rayleigh quotient. To solve (7), one can use the Gradient Ascent method, i.e., in the t-th iteration the vector v_t is updated as

$$v_t = v_{t-1} + \eta \nabla_v J(v_{t-1}),$$

where $\nabla_v J_v(v_{t-1}) \propto \hat{A} v_{t-1} - \frac{v_{t-1}^T \hat{A} v_{t-1}}{v_{t-1}^T \hat{B} v_{t-1}} \hat{B} v_{t-1}$ and η is the stepsize. Thus, to design DP methods, one natural approach is based on the idea of DP-SGD, which is a commonly used method for DP Empirical Risk Minimization (ERM) and DP Deep Learning [Abadi et al., 2016, Wu et al., 2017, Wang et al., 2018, Bassily et al., 2014]. The idea of DP-SGD is injecting some Gaussian noise into the (stochastic) gradient in each iteration. That is

$$v_t = v_{t-1} + \eta(\nabla_v J(v_{t-1}) + \zeta_{t-1}),$$

where ζ_{t-1} is a Gaussian vector where the variance of each coordinate is proportional to the sensitivity of $\nabla_v J(v_{t-1})$. However, the main challenge is that, unlike the objective functions in ERM or Deep Learning where the sensitivity of gradient is $O(\frac{1}{n})$, our objective function (Rayleigh quotient) cannot be decomposed into a sum of loss functions, which means the sensitivity of $\nabla_v J(v_{t-1})$ is larger and even could be unbounded. Thus, we cannot use DP-SGD based methods and need new approaches.

Based on the specific structure of $\nabla_v J_v(v_{t-1})$, here we propose a new method namely DP-Rayleigh Flow. Specifically, instead of injecting noise to the gradient, we add noises to matrices $\hat{A}$ and $\hat{B}$ in each iteration, see Algorithm 1 for details. However, compared with the original Rayleigh Flow method as mentioned above, we need

some modifications. First, instead of using some fixed stepsize η, in each iteration of Algorithm 1 we rescale it by $\rho_t = v_{t-1}^T \tilde{A}^t v_{t-1} / v_{t-1}^T \tilde{B}^t v_{t-1}$, where $\tilde{A}^t$ and $\tilde{B}^t$ are perturbed matrices in the t-th iteration, *i.e.,* we use $\frac{\eta}{\rho_t}$ as the stepsize, which is convenient for our following theoretical analysis. Secondly, after updating by using Gradient Ascent *i.e.,* calculating $C^t v_{t-1}$, in step 5 of Algorithm 1 we need to normalize the vector to ensure v_t has unit ℓ_2-norm. This step guarantees that the generalized Rayleigh quotient for the updated vector is at least as large as that of the initial vector. In the following we provide theoretical guarantees for our algorithm.

Algorithm 1 DP-Rayleigh Flow

1: **Input:** Matrices $\hat{A}$ and $\hat{B}$, initial parameter v_0 with $\|v_0\|_2 = 1$, step size η (will be specified later), iteration numbers m, privacy parameter ρ.
2: **for** $t = 1, \cdots, m$ **do**.
3: Denote $\tilde{A}^t = \hat{A} + Z_1^t$, $\tilde{B}^t = \hat{B} + Z_2^t$, where Z_1^t and Z_2^t are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2 m}{n^2 \rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2 m}{n^2 \rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
4: Denote $\rho_t = \frac{v_{t-1}^T \tilde{A}^t v_{t-1}}{v_{t-1}^T \tilde{B}^t v_{t-1}}$ and $C^t = I + \frac{\eta}{\rho_t}(\tilde{A}^t - \rho_t \tilde{B}^t)$
5: Update $v_t = \frac{C^t v_{t-1}}{||C^t v_{t-1}||_2}$.
6: **end for**
7: **return** v_m.

Theorem 2. Under Assumption 1, for any $\rho > 0$ Algorithm 1 is ρ-CDP.

Before showing the estimation error of the output in Algorithm 1, we first introduce several notations and assumptions. The following theorem indicates that when n is sufficiently large, the generalized eigenvalue of the perturbed matrices is close to the generalized eigenvalue of the underlying matrices.

Theorem 3. Let $\tilde{\lambda}_k^t$ be the kth generalized eigenvalues of $(\tilde{A}^t, \tilde{B}^t)$, where $(\tilde{A}^t, \tilde{B}^t)$ are the perturbed matrices in the t-th iteration. Under Assumption 2, given any failure probability $\zeta > 0$, let constants $0 \le b < \min_{j \in [d]} \frac{\lambda_j}{2\lambda_j^2 + 1}$, $0 \le c$ and if n is sufficiently large such that, $n \ge \tilde{\Omega}(\max\{\frac{d}{c^2 \lambda_{\min}^2(B)}, \frac{d}{b^2 \mathrm{cr}^2(A,B)}, \frac{\sqrt{dm \log \frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{dm \log \frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$. Then with probability at least $1 - \zeta$, there exists constants a such that for all $t \in [m]$,

$$(1-a)\lambda_j \le \tilde{\lambda}_j^t \le (1+a)\lambda_j,$$
$$(1-c)\lambda_j(B) \le \lambda_j(\tilde{B}^t) \le (1+c)\lambda_j(B) \quad (8)$$
$$C_{\text{lower}}\kappa(B) \le \kappa(\tilde{B}^t) \le C_{\text{upper}}\kappa(B) \quad (9)$$

where $C_{\text{lower}} = \frac{1-c}{1+c}$, $C_{\text{upper}} = \frac{1+c}{1-c}$ Furthermore, we have $\tilde{\lambda}_2^t \le \gamma \tilde{\lambda}_1^t$, where $\gamma = \frac{(1+a)\lambda_2}{(1-a)\lambda_1}$.

Theorem 4 (Informal)**.** Under Theorem 3 and choose the stepsize η such that $\eta \lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B) + \gamma}} < \frac{1}{2}.$$

Then if n is sufficient large, in Algorithm 3 we set $m = O(\log n)$ and if the input vector v_0 with $\|v_0\|_2 = 1$ satisfying $\frac{|\langle v^*, v_0\rangle|}{\|v^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$ with

$$\theta(A,B) = \min\{\frac{1}{8C_{\text{upper}}\kappa(B)}, \frac{1/\gamma - 1}{3C_{\text{upper}}\kappa(B)}, \frac{1-\gamma}{30(1+c)C_{\text{upper}}^2\eta\lambda_{\max}(B)\kappa^2(B)\{C_{\text{upper}}\kappa(B)+\gamma\}}\}, \quad (10)$$

we have the following with probability at least $1 - \zeta$,

$$1 - \frac{\langle v^*, v_m\rangle}{\|v^*\|_2} \le O(\frac{\theta(A,B)}{(1-\nu)^2} \times \big(\frac{1}{\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log d}{n} + \frac{1}{\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log n \log d \log\frac{1}{\zeta}}{n^2\rho}\big)), \quad (11)$$

where

$$\lambda_{gap} = \min_{j>1} \frac{\lambda_1 - (1+a)\lambda_j}{\sqrt{1+\lambda_1^2}\sqrt{1+(1-a)^2\lambda_j^2}} \quad (12)$$

is the eigengap for the GEP.

Similarly, for the deterministic setting where $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and we set some appropriate parameters in Algorithm 3, with probability at least $1 - \zeta$

$$1 - \frac{\langle \tilde{v}, v_m\rangle}{\|\tilde{v}\|_2} \le \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log d\log\frac{1}{\zeta}}{n^2\rho}). \quad (13)$$

Remark 1. Since $\theta(A,B), \lambda_{gap}, \mathrm{cr}(A,B)$ and v all only depend on the underlying matrices A and B. Thus the output could achieve an error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n^2\rho})$ under the stochastic setting and deterministic setting respectively (if we omit other terms). Note that in the non-private case, the optimal rate is $O(\frac{d}{n})$ for many statistical models such as PCA or CCA [Cai et al., 2013, Gao et al., 2015] if each $\|x_i\|_2 \le O(1)$. Thus, based on Theorem 1 we can see it is possible to obtain privacy nearly for free when $\rho > \frac{1}{n}$ in the statistical setting.

One major issue in Theorem 4 is we need to assume the initial vector v_0 is close enough to v^* such that $\frac{|\langle v^*, v_0\rangle|}{\|v^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$. In general, this condition is necessary since in general GEP is non-concave and the Gradient Ascent

method can only ensure the parameter converges to some local maximum. However, with some additional assumptions and n is the sufficiently large, in the following we show how to find such an initial vector privately and efficiently.

Note that in the non-private case, originally finding the K leading generalized eigenvectors for matrix pair $(\hat{A}, \hat{B})$ is equivalent to solve the following optimization problem:

$$\min_{U \in \mathbb{R}^{d\times K}} -\text{tr}(U^T \hat{A} U), \text{ s.t. } U^T \hat{B} U = I_K. \quad (14)$$

Due to the non-convexity of the previous problem, motivated by [Tan et al., 2018, Vu et al., 2013a, Wang and Xu, 2020] here we consider a convex relaxation with a LASSO penalty, i.e.,

$$\min_{P\in\mathbb{R}^{d\times d}} -\text{tr}(\hat{A}P) + \phi\|P\|_{1,1}, \\ \text{s.t. } \|\hat{B}^{\frac{1}{2}} P \hat{B}^{\frac{1}{2}}\|_{nu} \le K, \|\hat{B}^{\frac{1}{2}} P \hat{B}^{\frac{1}{2}}\|_2 \le 1, \quad (15)$$

where for a matrix A, $\|A\|_{nu}$ is defined as the sum of its singular values, $A^{\frac{1}{2}}$ is the square root of A, and $\|A\|_{1,1}$ is the ℓ_1-norm of the vector of row-wise ℓ_1 norm of A.

Our private estimator is based on (15). That is, instead of using the empirical matrices $\hat{A}$ and $\hat{B}$, we use their perturbed version to ensure DP. Specifically, we will solve the following optimization problem:

$$\hat{P} = \arg\min_{P\in\mathbb{R}^{d\times d}} -\text{tr}(\tilde{A}P) + \phi\|P\|_{1,1}, \\ \text{s.t. } \|\tilde{B}^{\frac{1}{2}} P \tilde{B}^{\frac{1}{2}}\|_{nu} \le K, \|\tilde{B}^{\frac{1}{2}} P \tilde{B}^{\frac{1}{2}}\|_2 \le 1, \quad (16)$$

where $\tilde{A} = \hat{A} + Z_1$, $\tilde{B} = \hat{B} + Z_2$ and Z_1 and Z_2 are symmetric Gaussian matrices to ensure DP.

Since the optimization problem (16) is convex, we can follow the approach in [Wang and Xu, 2020] to solve it by using ADMM method (see Algorithm 2 for the details).

Informally we have the following result.

Theorem 5 (Informal)**.** Under Assumption 1, the solution of the optimization problem (16) is ρ-CDP. Moreover, under Assumption 2 and assume that $\|E_A\|_{\infty,\infty}, \|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, and n is sufficiently large, take $\phi = \tilde{O}(\lambda_{\max}(B)\lambda_1(\frac{\sqrt{d}}{\sqrt{n}} + \frac{\sqrt{d}}{n\sqrt{\rho}}))$, $K = 1$ in (16). Then the largest eigenvalue of the matrix $\hat{P}$ which is denoted as v_0, satisfies $\langle v_0, v^*\rangle \ge 1 - \theta(A,B)/2$ with high probability. For a matrix A, $\|A\|_{\infty,\infty}$ is defined as the maximal absolute value among the entries in A.

In the above theorem we need to assume that $\|E_A\|_{\infty,\infty}$, $\|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, these assumptions hold in the deterministic setting where $E_A = E_B = 0$. In the stochastic setting, we can show these assumptions hold for PCA, CCA and SIR if $\|x_i\|_2 \le 1$ (see the Proof of Theorem 1).

Algorithm 2 Privately Finding an Initial Vector

1: **Input:** Matrices $\hat{A}$ and $\hat{B}$, privacy parameters ρ, tuning parameter ϕ, ADMM parameter v, and convergence criterion β.
2: Initialize matrices P_0, H_0 and Γ_0. Set $t = 0$
3: Let $\tilde{A} = \hat{A} + Z_1$, $\tilde{B} = \hat{B} + Z_2$ and Z_1 and Z_2 are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2}{2n^2\rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2}{2n^2\rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
4: Update P by solving the following lasso problem:

$$P_{t+1} = \arg\min \frac{v}{2}\|\tilde{B}^{\frac{1}{2}} P \tilde{B}^{\frac{1}{2}} - H_t + \Gamma_t\|_F^2 - \text{tr}(\tilde{A}P) + \phi\|P\|_{1,1}.$$

5: Let $\sum_{i=1}^d w_j a_j a_j^T$ be the singular value decomposition of $\Gamma_t + \tilde{B}^{\frac{1}{2}} P_{t+1} \tilde{B}^{\frac{1}{2}}$ and let

$$\gamma^* = \arg\min_{\gamma>0} \gamma, \\ \text{s.t. } \sum_{j=1}^d \min\{1, \max\{w_j - \gamma, 0\}\} \le K.$$

Update H by $H_{t+1} = \sum_{j=1}^d \min\{1, \max\{w_j - \gamma^*, 0\} a_j a_j^T$.
6: Update Γ as $\Gamma_{t+1} = \Gamma_t + \tilde{B}^{\frac{1}{2}} P_{t+1} \tilde{B}^{\frac{1}{2}} - H_{t+1}$.
7: If $\|P_{t+1} - P_t\|_F > \beta$, let $t = t + 1$ and repeat the procedure 4-6.
8: **return** The leading eigenvector of P_{t+1}.

5 HIGH DIMENSIONAL SPARSE CASE

In the previous section, we showed the upper bounds of the estimation error in stochastic and deterministic settings. However, in the high dimensional case where $d \gg n$, the previous two bounds will be quite large so that their rates become trivial. To address the high dimensionality issue, in this section we consider the sparse GEP instead, *i.e.,* problem (3) and (4). Specifically, we propose a truncated version of Algorithm 1, namely DP-Truncated Rayleigh Flow, see Algorithm 3 for details. Compared with Algorithm 1, there is an additional truncation step. That is, we select the indices with largest k magnitude of the vector, keep the entries of vectors among these indices and let the remain entries be zero. Intuitively, the truncation step could project the vector onto a low dimensional space (and thus the effective dimension now becomes to k instead of d), and it will diminish the noises we added to $\hat{A}$ and $\hat{B}$. Note that the idea of truncating the vector to enforce it be sparse has also been used in other DP machine learning problems, such as [Cai et al., 2019,

Algorithm 3 DP-Truncated Rayleigh Flow

Input: Matrices $\hat{A}$ and $\hat{B}$, sparsity k, initial parameter v_0 is a k-sparse vector with $\|v_0\|_2 = 1$, step size η, iteration number m, privacy parameter ρ.
for $t = 1, \cdots, m$ **do**.
  Denote $\tilde{A}^t = \hat{A} + Z_1^t$, $\tilde{B}^t = \hat{B} + Z_2^t$, where Z_1^t and Z_2^t are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2 m}{n^2\rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2 m}{n^2\rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
  Denote $\rho_t = \frac{v_{t-1}^T \tilde{A}^t v_{t-1}}{v_{t-1}^T \tilde{B}^t v_{t-1}}$ and $C^t = I + (\eta/\rho_t)(\tilde{A}^t - \rho_t \tilde{B}^t)$.
  Update $v_t' = \frac{C^t v_{t-1}}{||C^t v_{t-1}||_2}$.
  Let $F_t = \text{supp}(v_t', k)$ be the set of indices of v_t' with the largest k absolute values.
  Denote $\hat{v}_t = \text{truncate}(v_t', F_t)$, *i.e.,* $\hat{v}_t$ is the truncated vector of v_t' by setting $(v_t')_i = 0$ if $i \notin F_t$.
  Update $v_t = \frac{\hat{v}_t}{\|\hat{v}_t\|_2}$.
end for
return v_m.

Wang et al., 2019, Wang and Gu, 2020, Hu et al., 2021] for DP-ERM and [Ge et al., 2018] for DP-Sparse PCA. However, as we mentioned, unlike those objective functions, the Rayleigh quotient cannot be decomposed as a sum of functions, it is unknown whether truncation step is indeed helpful. We will provide an affirmative answer in this section.

Theorem 6. Under Assumption 1, for any $0 < \rho$, Algorithm 3 is ρ-CDP.

Before providing the estimation error of Algorithm 3 we first provide the following notations.

Notations: For v_s^* in (4) we denote $V = \text{supp}(v_s^*)$ as the index set corresponding to the non-zero elements of v_s^*. Let $F \subseteq [d]$ be a superset of V with $|F| = k'$, where $k' = 2k + s$ and k is in Algorithm 3. Let $\lambda_j(F)$, $\tilde{\lambda}_k^t(F)$ and $\hat{\lambda}_j(F)$ be the j-th generalized eigenvalue of the matrix pairs (A_F, B_F), $(\tilde{A}_F^t, \tilde{B}_F^t)$ and $(\hat{A}_F, \hat{B}_F)$, respectively. Denote $\text{cr}(k') = \inf_{F:|F|\le k'} \text{cr}(A_F, B_F)$.

Similar to Theorem 3, we first show that when n is sufficiently large, then the generalized eigenvalue (restricted to the set F) of the perturbed matrices is close to the generalized eigenvalue of the underlying matrices.

Theorem 7. Under Assumption 2, given any failure probability $\zeta > 0$, if n is sufficiently large such that $n \ge \Omega(\max\{\frac{k'}{b^2\text{cr}^2(k')}, \frac{k'}{c^2\lambda_{\min}^2(B)}, \frac{\sqrt{k'm\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{k'm\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$ for some constants $c \ge 0$ and $0 \le b < \min_{j\in[d]} \frac{\lambda_j(F)}{2\lambda_j^2(F)+1}$. Then with probability at least $1-\zeta$, there exists constants a and c such that for all $t \in [m]$,

$$(1-a)\lambda_j(F) \le \tilde{\lambda}_j^t(F) \le (1+a)\lambda_j(F), \quad (17)$$

$$(1-c)\lambda_j(B_F) \le \lambda_j(\tilde{B}_F^t) \le (1+c)\lambda_j(B_F), \quad (18)$$

$$C_{\text{lower}}\kappa(B) \le \kappa(\tilde{B}_F^t) \le C_{\text{upper}}\kappa(B) \quad (19)$$

where $C_{\text{lower}} = \frac{1-c}{1+c}$, $C_{\text{upper}} = \frac{1+c}{1-c}$. Furthermore, we have

$$\tilde{\lambda}_2^t(F) \le \gamma\tilde{\lambda}_1^t(F), \quad (20)$$

where $\gamma = \frac{(1+a)\lambda_2(F)}{(1-a)\lambda_1(F)}$.

In the following we provide the statistical error of our private estimator if n is sufficiently large and the initial vector is close the optimal solution with $m = O(\log n)$.

Theorem 8 (Informal)**.** Under Theorem 7 with $k' = 2k+s$ and choose $k = Cs$ for sufficiently large C. In addition, choose stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 + 2\sqrt{\frac{s}{k}} + 2\frac{s}{k}} \times \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B)+\gamma}} < \frac{1}{2}.$$

Then if n is sufficiently large, we set $m = O(\log n)$ in Algorithm 3. We have the following with probability at least $1-\zeta$ if the input k-sparse vector v_0 with $\|v_0\|_2$ satisfying $\frac{|\langle v_s^*, v_0\rangle|}{\|v_s^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$ with $\theta(A,B)$ given in (10).

$$1 - \frac{\langle v_s^*, v_m\rangle}{\|v_s^*\|_2} \le O(\frac{\theta(A,B)}{(1-\nu)^2}(\frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d}{n} + \frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log n\log d\log\frac{1}{\zeta}}{n^2\rho})). \quad (21)$$

Similarly, for the deterministic setting where $E_A = E_B = 0$ and $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and with some additional mild assumptions. If we set some appropriate parameters in Algorithm 3, with probability at least $1-\zeta$

$$1 - \frac{\langle \tilde{v}_s, v_t\rangle}{\|\tilde{v}_s\|_2} \le \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d\log\frac{1}{\zeta}}{n^2\rho}). \quad (22)$$

From Theorem 8 we can find that, the error in the deterministic setting is $\tilde{O}(\frac{s\log d}{n^2\rho})$, while the statistical error of Algorithm 3 will be $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ (if we omit other terms). These two bounds only depend on logarithmic of d instead of polynomial in the low dimensional case. Moreover, the same as in the low dimensional case, we can obtain privacy for free in the statistical setting.

Corollary 1. If we transform the above upper bounds in CDP to (ϵ, δ)-DP via Lemma 1, we can see for PCA under deterministic setting, the output of Algorithm 1 could

achieve an error of $\tilde{O}(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon})$, which is near optimal [Dwork et al., 2014]. For sparse PCA under the stochastic setting where A is the covariance matrix and $B = I$, if we further assume that $\|x\|_2 \leq 1$. Then the output of Algorithm 3 could achieve a statistical error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d\log\frac{1}{\delta}}{n^2\epsilon^2})$. [Wang et al., 2019] provides the first result on the problem in the local DP model, instead of the central model. Specifically, it shows that the near optimal statistical rate is $\tilde{O}(\frac{s\log d\log\frac{1}{\delta}}{n\epsilon^2})$ under stochastic setting. Compared with our results, we can see a gap between the central and the local model for sparse PCA.

Corollary 2. For the problem of sparse CCA under the stochastic setting and $\|x_2\|_2 \leq 1$. The output of Algorithm 3 could achieve an error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$, where $d = d_1 + d_2$. Under deterministic setting, the output of Algorithm 3 could achieve an error of $\tilde{O}(\frac{s\log d}{n^2\rho})$. In the low dimension setting, the error will be $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n^2\rho})$, respectively. Moreover, we have similar results for SIR if $\|x\|_2 \leq 1$. Note that these are the first results on the estimation error for CCA and SIR in the DP model.

Corollary 3. For FDA, the output of Algorithm 1 and 3 could achieve an error of $\tilde{O}(\frac{d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n^2\rho})$ in the low dimension and high dimensional sparse case respectively if $\|x\|_2 \leq 1$. To our best knowledge, this is the first theoretical result for FDA in the DP model.

Similar to the low dimension case, here we still need a good initialization v_0. However, unlike the low dimension case, here we cannot use Algorithm 2 to find such a initialization due to the assumption of $d \gg n$. Thus, we leave it as an open problem for privately finding such a initialization. Fortunately, in experiments we find randomly sample an initial vector can already achieve good performance.

Experimental studies: In Appendix, we provide empirical studies on the behaviors of our methods for (sparse) PCA, CCA and FDA on several real-world and synthetic data.

6 LOWER BOUNDS OF DP-GEP

In previous sections, we showed that for GEP in the CDP model under Assumption 1 and 2, it is possible to achieve an error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ in low and high dimension sparse case under the statistical setting respectively. However, there are several questions left. First, can we further improve the error, i.e., what is the lower bound of error for GEP in the CDP model? Secondly, since all of our previous results are for the CDP or (ϵ, δ)-DP model. Thus, our question is, can we achieve similar results in the ϵ-DP model? In this section, we first show that the previous methods are near optimal for (sparse) PCA and (sparse) SIR in the CDP model. For the second one, we provide negative results by showing lower bounds of (sparse) PCA under the stochastic setting in ϵ-DP. Specifically, we show the following results.

Theorem 9 (Lower Bounds for Low Dimensional PCA)**.** For $0 < \epsilon \leq 1$, if $n \geq \Omega(\frac{d}{\epsilon})$, then for any ϵ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d^2}{n^2\epsilon^2}). \tag{23}$$

Moreover, for any $\rho > 1$, if $n \geq \Omega(\max\{d, \frac{\sqrt{d}}{\sqrt{\rho}}\})$, then for any ρ-CDP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d}{n^2\rho}). \tag{24}$$

Here v^* is the leading eigenvector of $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]]$.

Theorem 10 (Lower Bounds for High Dimensional Sparse PCA)**.** For $0 < \epsilon \leq 1$, if $n \geq \Omega(\frac{s\log d}{\epsilon})$, then for any ϵ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$, if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and its largest eigenvector v^* of the covariance matrix $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]$ is s-sparse, such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{(s\log d)^2}{n^2\epsilon^2}). \tag{25}$$

Moreover, for any $\rho > 0$, if n is sufficiently large such that $n \geq \Omega(\max\{s\log d, \frac{\sqrt{s\log d}}{\sqrt{\rho}}\})$, then for any ρ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$, if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and its largest eigenvector v^* of the covariance matrix $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]$ is s-sparse, and

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho}). \tag{26}$$

Next we consider the lower bounds of SIR in the CDP model, for simplicity we only consider the case where $k = 2$. That is we have two classes $Y = 1$ and $Y = 2$.

Theorem 11. For any $\rho > 0$, if $n \geq \Omega(\max\{d, \frac{\sqrt{d}}{\sqrt{\rho}}\})$, then for any ρ-CDP algorithm with output v_{priv}, there exists an instance $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d}{n^2\rho}). \tag{27}$$

Here $\|v^*\|_2 = 1$ is the leading generalized eigenvector of the corresponding SIR.

Theorem 12. For any $\rho > 0$, if n is sufficiently large such that $n \geq \Omega(\max\{s\log d, \frac{\sqrt{s\log d}}{\sqrt{\rho}}\})$, then for any ρ-CDP

algorithm with output v_{priv}, there exists an instance $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho}). \quad (28)$$

Here $\|v^*\|_2 = 1$ is the leading generalized eigenvector of the corresponding sparse SIR with $\|v^*\|_0 \leq s$.

7 CONCLUSIONS

In this paper we provided the first study on the theoretical behaviors of the (sparse) Generalized Eigenvalue Problem (GEP) in the Differential Privacy (DP) model. Specifically, we considered both stochastic setting and deterministic setting in the low dimensional and high dimensional sparse cases. With some additional assumptions, we showed that our algorithms could achieve near optimal rates of error under the stochastic setting in both low dimensional and high dimensional sparse cases. Moreover, we provided the lower bound of (sparse) GEP in the ϵ-DP model to show a gap of the problem in the (ϵ, δ)-DP model.

However, there are still several unsolved problems left. First, from lower bounds and upper bounds of the error we can see that there is still a gap of $\text{Poly}(\log n)$ factor. Thus, can we further improve the upper bounds of error? Secondly, in the low dimension case, we discussed how to find an appropriate initial vector privately and efficiently. However, our approach cannot be extended to the high dimensional sparse case since we need to assume the sample size is large enough such that $n \gg d$, which violates the high dimension assumption. Thus, how do we find the initial vector privately in this case? Thirdly, for the lower bounds we proposed, we only considered the case for (sparse) PCA with sub-Gaussian distribution, where $\rho(E_A, k), \rho(E_B, k) = O(\sqrt{\frac{k\log d}{n}})$ and $\|E_A\|_2, \|E_B\|_2 = O(\sqrt{\frac{d}{n}})$. Thus, our question is, can we provide more general lower bounds which involve general $\rho(E_A, k)$ and $\rho(E_B, k)$? Finally, in the lower bound part we mainly focused on the stochastic setting. In the deterministic setting, [Dwork et al., 2014] provided the lower bound of PCA in the low dimension case. However, the lower bound of sparse PCA is still unknown. We will leave these open problems as future work.

ACKNOWLEDGEMENTS

Lijie Hu, Zihang Xiang and Di Wang are supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Di Wang was also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

[Acharya et al., 2021] Acharya, J., Sun, Z., and Zhang, H. (2021). Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR.

[Balcan et al., 2016] Balcan, M.-F., Du, S. S., Wang, Y., and Yu, A. W. (2016). An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309.

[Barber and Duchi, 2014] Barber, R. F. and Duchi, J. C. (2014). Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*.

[Bassily et al., 2014] Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE.

[Biswas et al., 2020] Biswas, S., Dong, Y., Kamath, G., and Ullman, J. (2020). Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*.

[Blum et al., 2005] Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005). Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138. ACM.

[Bun and Steinke, 2016] Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer.

[Cai et al., 2013] Cai, T. T., Ma, Z., Wu, Y., et al. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

[Cai et al., 2019] Cai, T. T., Wang, Y., and Zhang, L. (2019). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*.

[Chaudhuri et al., 2012] Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. *Advances in Neural Information Processing Systems*, 25:989–997.

[Chaudhuri et al., 2013] Chaudhuri, K., Sarwate, A. D., and Sinha, K. (2013). A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14(1):2905–2943.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

[Dwork et al., 2014] Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. (2014). Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20. ACM.

[Gao et al., 2015] Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5):2168–2197.

[Ge et al., 2018] Ge, J., Wang, Z., Wang, M., and Liu, H. (2018). Minimax-optimal privacy-preserving sparse pca in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, pages 1589–1598.

[Golub and Van Loan, 1996] Golub, G. H. and Van Loan, C. F. (1996). Matrix computations. johns hopkins studies in the mathematical sciences.

[Gonem and Gilad-Bachrach, 2018] Gonem, A. and Gilad-Bachrach, R. (2018). Smooth sensitivity based approach for differentially private pca. In *Algorithmic Learning Theory*, pages 438–450.

[Hardt and Price, 2014] Hardt, M. and Price, E. (2014). The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869.

[Hardt and Roth, 2013] Hardt, M. and Roth, A. (2013). Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340. ACM.

[Hu et al., 2021] Hu, L., Ni, S., Xiao, H., and Wang, D. (2021). High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*.

[Imtiaz and Sarwate, 2017] Imtiaz, H. and Sarwate, A. D. (2017). Differentially-private canonical correlation analysis. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 283–287. IEEE.

[Imtiaz and Sarwate, 2019] Imtiaz, H. and Sarwate, A. D. (2019). Distributed differentially-private canonical correlation analysis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3112–3116. IEEE.

[Jin et al., 2019] Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*.

[Kamath et al., 2021] Kamath, G., Liu, X., and Zhang, H. (2021). Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*.

[Kapralov and Talwar, 2013] Kapralov, M. and Talwar, K. (2013). On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1395–1414. SIAM.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

[Li, 1991] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

[Liu and Altman, 2015] Liu, T. and Altman, R. B. (2015). Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *Journal of chemical information and modeling*, 55(7):1483–1494.

[Liu et al., 2022] Liu, X., Kong, W., Jain, P., and Oh, S. (2022). Dp-pca: Statistically optimal and differentially private pca. *arXiv preprint arXiv:2205.13709*.

[Massart, 2007] Massart, P. (2007). Concentration inequalities and model selection.

[Parkhomenko et al., 2009] Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1).

[Parlett, 1998] Parlett, B. N. (1998). *The symmetric eigenvalue problem*. SIAM.

[Shen, 2020] Shen, Y. (2020). Differentially private nonlinear canonical correlation analysis. In *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE.

[Stewart, 1979] Stewart, G. (1979). Pertubation bounds for the definite generalized eigenvalue problem. *Linear algebra and its applications*, 23:69–85.

[Strickert et al., 2009] Strickert, M., Keilwagen, J., Schleif, F.-M., Villmann, T., and Biehl, M. (2009). Matrix metric adaptation linear discriminant analysis of biomedical data. In *International Work-Conference on Artificial Neural Networks*, pages 933–940. Springer.

[Szarek, 1982] Szarek, S. J. (1982). Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, volume 169, page 185.

[Tan et al., 2018] Tan, K. M., Wang, Z., Liu, H., and Zhang, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1057–1086.

[Vershynin, 2009] Vershynin, R. (2009). On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE.

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

[Vu et al., 2013a] Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013a). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. *Advances in neural information processing systems*, 26.

[Vu et al., 2013b] Vu, V. Q., Lei, J., et al. (2013b). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947.

[Wang et al., 2019] Wang, D., Chen, C., and Xu, J. (2019). Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535. PMLR.

[Wang and Xu, 2020] Wang, D. and Xu, J. (2020). Principal component analysis in the local differential privacy model. *Theoretical computer science*, 809:296–312.

[Wang et al., 2018] Wang, D., Ye, M., and Xu, J. (2018). Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*.

[Wang and Gu, 2020] Wang, L. and Gu, Q. (2020). A knowledge transfer framework for differentially private sparse learning. In *AAAI*, pages 6235–6242.

[Wu et al., 2017] Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322.

[Yuan et al., 2019] Yuan, G., Shen, L., and Zheng, W.-S. (2019). A decomposition algorithm for the sparse generalized eigenvalue problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6113–6122.

Privacy-preserving Sparse Generalized Eigenvalue Problem: Supplementary Materials

1 Experiments

In this section, we will investigate the practical performance of our previous methods to (sparse) PCA, (sparse) CCA and (sparse) FDA over both synthetic data and real-world datasets.

Datasets: We conduct our algorithms on four datasets from the UCI Machine Learning Repository [Dua and Graff, 2017] that have been studied before such as [Yuan et al., 2019]: 1) Covertype, a 7-label classification dataset which contains 581,012 data samples, each sample is a 54 dimension vector; 2) Dota2, a binary classification dataset which contains 92,650 data samples, and each sample is a 116 dimension vector; 3) Optical, a 7-label classification dataset which contains 325,834 data samples, each sample is a 174 dimension vector; 4) Synthetic dataset. Each data point in this dataset is a vector and each entry is sampled from the standard normal distribution. For this dataset, the data size and dimension are manually set and are subject to specific experiment settings. For the real world datasets, we stick to the original labels, for the synthetic dataset, we randomly assign 0 or 1 (for binary classification in FDA) to each data point. To ensure each data sample satisfies $\|x_i\|_2 \leq 1$ in this paper, we normalize each data sample by the maximal ℓ_2 norm of all data samples within the dataset.

Experiment setup: For each statistical model (PCA, CCA, FDA) we conduct three experiments on each dataset to show how the error varies with respect to different factors: 1) In both low dimension (LD) and and high dimensional sparse (HDS) cases, we consider the estimation error with different size of training data. Note that in the HDS case, we also fix the underlying sparsity s to be 30; 2) Moreover, in the HDS case, we also study different sparsity of the underlying parameter. In the LD case, the number of dimension we use is $d = 20$ for PCA, $d = 30$ for CCA and $d = 10$ for FDA, which correspond to the first d features in the original data. To see how the size of the data n (which corresponds the selected first n data samples in the original dataset) impacts the error, we set various data sizes for each dataset. See Table 1 for details.

	Covertype	Dota2	Optical	Synthetic
CCA	1.5k, 2.9k, 4.3k, 5.7k	3k, 5k, 7k, 9k	9k, 17k, 26k, 32k	1k, 1.5k, 2k, 3k
PCA	150k, 290k, 430k, 570k	30k, 50k, 70k, 90k	90k, 170k, 260k, 320k	30k, 90k, 270k, 810k
FDA	150k, 290k, 430k, 570k	30k, 50k, 70k, 90k	90k, 170k, 260k, 320k	30k, 90k, 270k, 810k

Table 1: Different data size setup for each dataset in LD case and HDS case with fixed sparsity 30.

To be consistent with the previous empirical studies on DP-PCA or DP-CCA such as [Wang et al., 2019]. We will consider (ϵ, δ)-DP model. That is for given (ϵ, δ), we run our algorithm with $\rho = (\sqrt{\epsilon + \log 1/\delta} - \sqrt{\log 1/\delta})^2$. For all experiments, we test different privacy levels: $\epsilon = \{2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ with $\delta = \frac{1}{n^{1.1}}$. For baseline methods, as we mentioned earlier, there is no previous results on DP-GEP. Thus, we only implement and compare our algorithms. For the initial vector, although theoretically we need assume it is close to the optimal solution. However, as we found in the experiments, random sampling an initial vector could already achieve good performance even without using Algorithm 2. Thus, we random sample an initial vector for convenience.

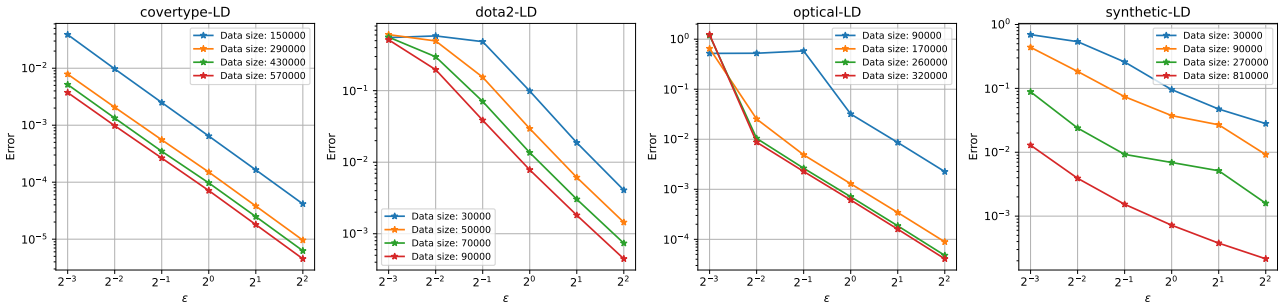
Hyperparameters: For each experiment, we set the number of iterations T as 15 for convenience. Note that this is reasonable since its has the same scale with $\log n$. We also set the step size $\eta = 0.1$ for CCA, and $\eta = 1$ for PCA and FDA. For CCA, in each experiment, we set $d_1 = d_2 = \hat{D}/2$ which means that $x_i, y_i \in \mathbb{R}^{\hat{D}/2}$ where $\hat{D}$ is the dimension of the data sample, x_i and y_i is the first and second half part of that data sample respectively. For FDA, we only consider the binary classification task.

Error metric: For all experiments we compute the error as follows. First we run the best known method in [Tan et al., 2018] for (sparse) GEP in the non-private case to get the optimal solution, which is denoted as v^*. Then we run our previous algorithms to get private estimator v_{priv}. Finally we compute the similarity of v_{priv} and v^* according to Definition 5.

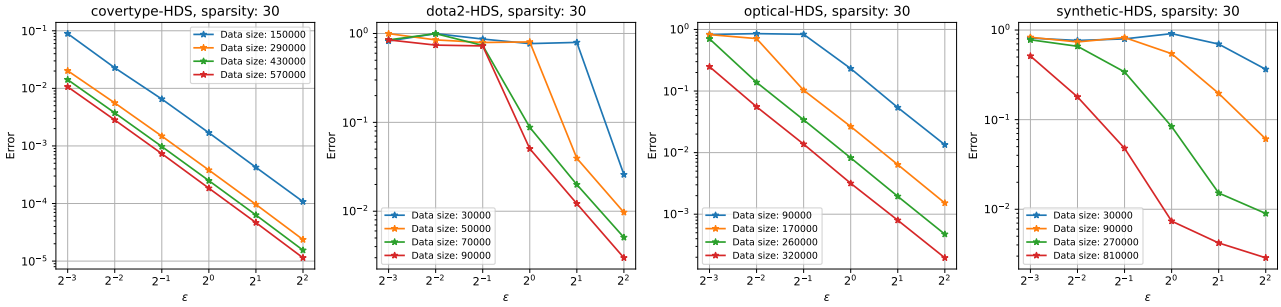
Experimental Results Figure 1 to Figure 3 are the results for PCA, CCA and FDA respectively. We can see that

- For all the models, the error will decrease when the data size becomes larger. However, there are still some exceptions such as the Covertype, synthetic data and Dota2 for CCA in the case where ϵ is large. Fortunately, since the error is in the scale of 10^{-4}, these deviations are acceptable due to other errors.
- From the last row of these Figures we can see that when the underlying sparsity is smaller, the error will decrease with fixed data size and small privacy level ϵ. This is due to that the estimation error depends on $\sqrt{s}$, which has been showed theoretically in the previous section. However, we can also find that with different underlying sparsity, the error does not change too much, which is due to that the size of the dataset is far greater than the sparsity and the error bound depends on $O(\frac{s}{n^2\epsilon^2})$ theoretically.
- From all experiments we can also find that when ϵ becomes larger (which implies ρ is larger) the error tend to decrease (although there are some exceptions, as we mentioned earlier they are still acceptable due to the scale of the error). Moreover, we can also see that in CCA, the error tends to be stable when ϵ is larger. We conjecture this is due to that the error of $O(\frac{d}{n})$ and $O(\frac{s\log d}{n})$ in Theorem 4 and Theorem 8 dominates the error caused by privacy. Also these errors only depend on the underlying distribution of the data and our datasets.

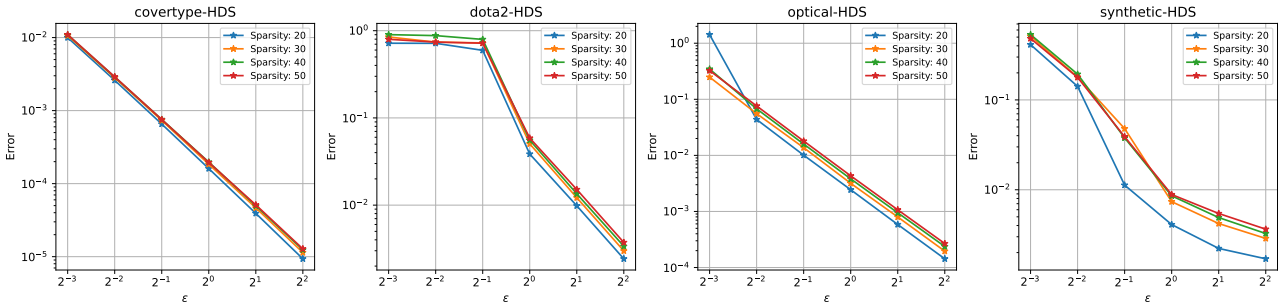
In total, all the previous experimental results support our previous theoretical analysis.



(a) Low dimension(LD) case with varying data size

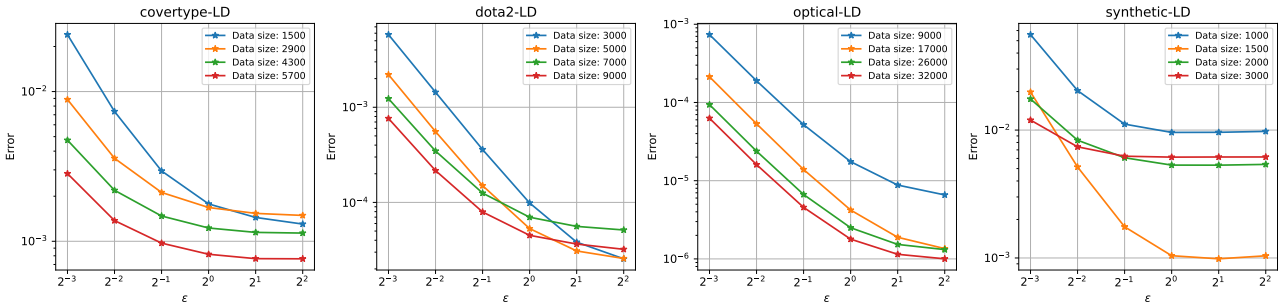


(b) High dimension sparse(HDS) case with varying data size(sparsity=30)

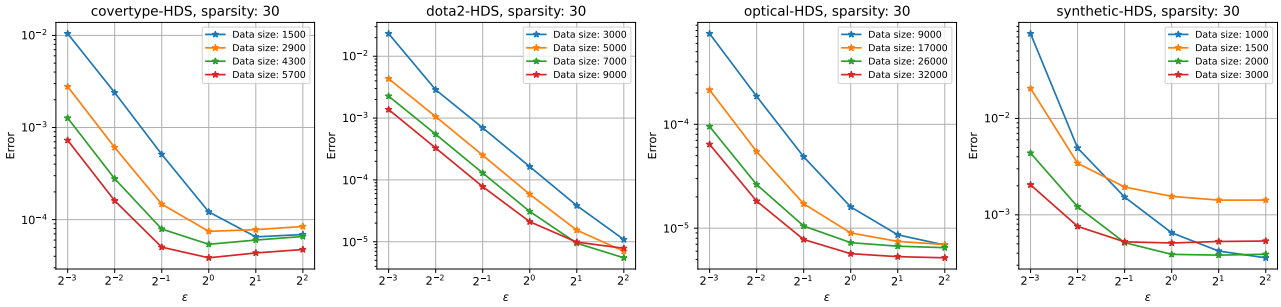


(c) High dimension sparse(HDS) case with varying spasity specification

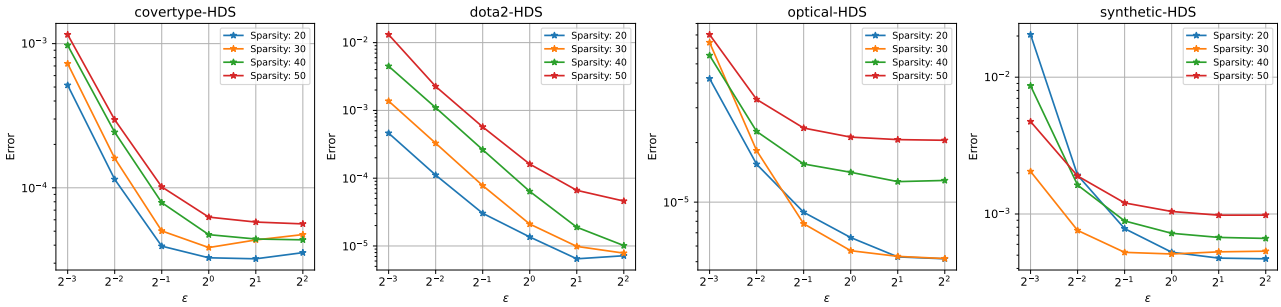
Figure 1: PCA application



(a) Low dimension(LD) case with varying data size

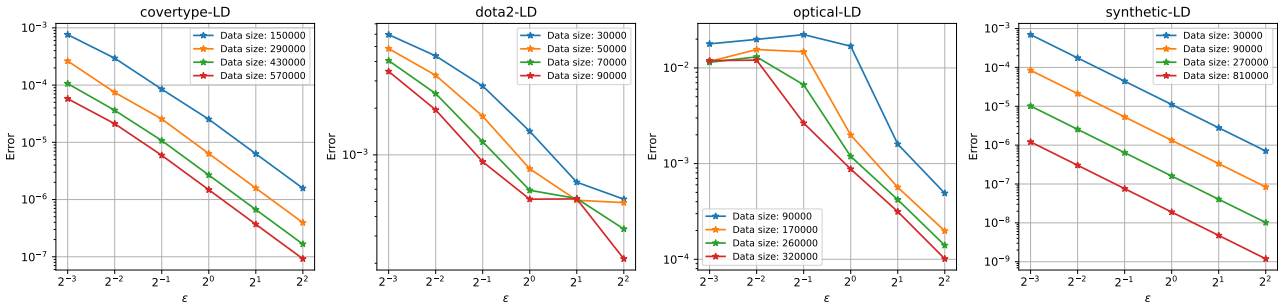


(b) High dimension sparse(HDS) case with varying data size(sparsity=30)

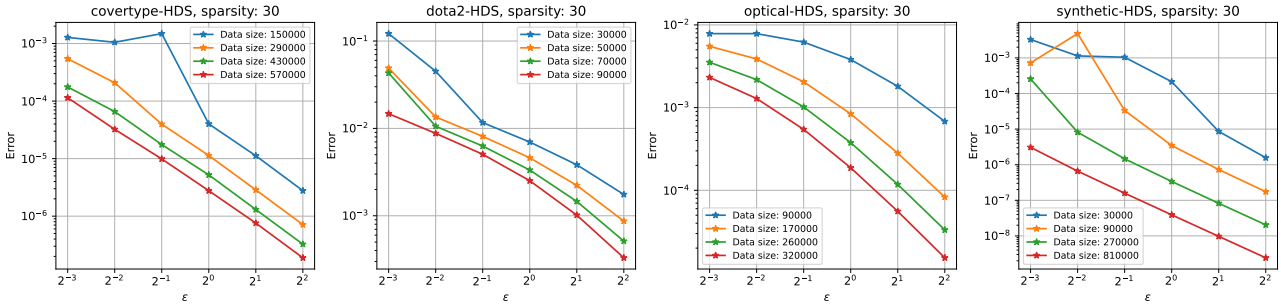


(c) High dimension sparse(HDS) case with varying spasity specification

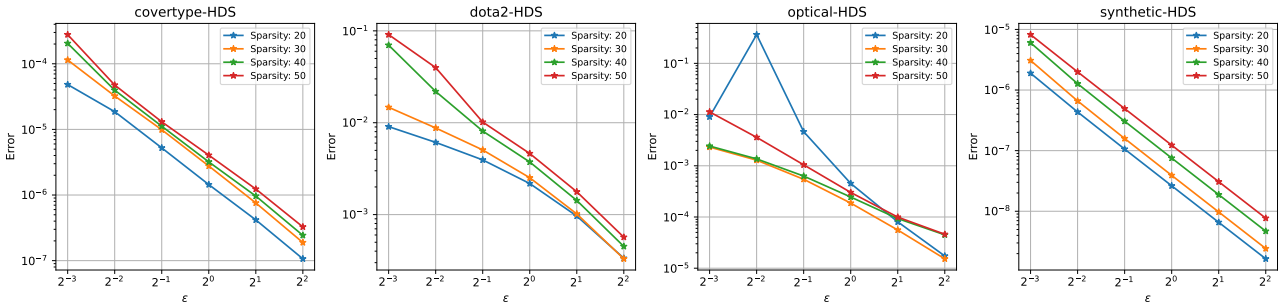
Figure 2: CCA application



(a) Low dimension(LD) case with varying data size



(b) High dimension sparse(HDS) case with varying data size(sparsity=30)



(c) High dimension sparse(HDS) case with varying spasity specification

Figure 3: FDA application

2 Technical Lemmas

Definition 6 (ϵ-Net)**.** Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\epsilon > 0$. A subset $\mathcal{S} \subseteq K$ is called an ϵ-net of K if every point in K is within a distance ϵ of some points of $\mathcal{S}$, *i.e.,*

$$\forall x \in k, \exists x_0 \in \mathcal{N} : d(x, x_0) \leq \epsilon.$$

The smallest possible cardinality of an ϵ-net of K is called the covering number of K and is denoted by $\mathcal{N}(K, d, \epsilon)$. Equivalently, covering number is the smallest number of closed balls with centers in K and radii ϵ whose union covers K.

Lemma 3 ([Vershynin, 2009])**.** Consider $S = \{u : \|u\|_2 \leq 1, \|u\|_0 \leq s\}$ be the set of all s-sparse unit vectors and let $\mathcal{S}_\epsilon$ be the ϵ-net of S then its covering number satisfies $\mathcal{N}(S, \tau) \leq e^s \cdot \frac{cd}{s\epsilon}$ with a universal constant c.

3 Proof of Theorem 1

Proof of Theorem 1. For the sensitivity of PCA, it has already been proven in some previous work on DP-PCA, see [Dwork et al., 2014] for details. We then consider **FDA**.

Assume that the samples can be divided in K classes. The $\mathcal{D}$ and $\mathcal{D}'$ are only different by deleting one data record (x_j, y_j) in the first class. For convenience, we denote $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ and $\mathcal{D}' = \{x_i', y_i'\}_{i=1, i\neq j}^n$, note that $\{(x_i', y_i') = (x_i, y_i)\}$. Then, the data sensitivity of $\hat{A}$ and $\hat{B}$ are as follows.

$$\begin{aligned}
&||\hat{B}(D) - \hat{B}(D')||_F \\
&= ||\frac{1}{n}\sum_{k=1}^{K} n_k \hat{\mu_k}\hat{\mu_k}^T - \frac{1}{n-1}\sum_{k=1}^{K} n_k' \hat{\mu_k}'\hat{\mu_k}'^T||_F \\
&= ||\frac{1}{n} n_1 \hat{\mu_1}\hat{\mu_1}^T - \frac{1}{n-1}(n_1 - 1)\hat{\mu_1}'\hat{\mu_1}'^T||_F \\
&= ||\frac{1}{n}\frac{1}{n_1}(\sum_{i\in C_1} x_i)(\sum_{i\in C_1} x_i)^T - \frac{1}{n-1}\frac{1}{n_1 - 1}(\sum_{i\in C_1} x_i')(\sum_{i\in C_1} x_i')^T||_F \\
&= ||\frac{1}{n}\frac{1}{n_1}(\sum_{i\in C_1, i\neq j} x_i + x_j)(\sum_{i\in C_1, i\neq j} x_i + x_j)^T - \frac{1}{n-1}\frac{1}{n_1 - 1}(\sum_{i\in C_1} x_i)(\sum_{i\in C_1} x_i)^T||_F \\
&\leq ||(\frac{1}{n}\frac{1}{n_1} - \frac{1}{n-1}\frac{1}{n_1 - 1})(\sum_{i\in C_1, i\neq j} x_i)(\sum_{i\in C_1, i\neq j} x_i)^T||_F + +||\frac{1}{n}\frac{1}{n_1} x_j x_j^T||_F \\
&\quad ||\frac{1}{n}\frac{1}{n_1}(\sum_{i\in C_1, i\neq j} x_i)x_j^T||_F + ||\frac{1}{n}\frac{1}{n_1} x_j(\sum_{i\in C_1, i\neq j} x_i)^T||_F \\
&\leq \frac{(n-1) + n_1}{n n_1 (n-1)(n_1 - 1)}(n_1 - 1)^2 + \frac{1}{n}\frac{1}{n_1}(n_1 - 1) + \frac{1}{n}\frac{1}{n_1}(n_1 - 1) + \frac{1}{n}\frac{1}{n_1} \leq \frac{4}{n}
\end{aligned} \tag{29}$$

Thus we can see the sensitivity of $\hat{B}(D)$ is upper bounded by $\frac{4}{n}$.

$$
\begin{aligned}
&||\hat{A}(D) - \hat{A}(D')||_F \\
&= ||\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i - \hat{u_k})(x_i - \hat{u_k})^T - \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i' - \hat{u_k}')(x_i' - \hat{u_k}')^T||_F \\
&\leq ||\frac{1}{n}\sum_{i\in C_1}(x_i - \hat{u_1})(x_i - \hat{u_1})^T - \frac{1}{n-1}\sum_{i\in C_1}(x_i' - \hat{u_1}')(x_i' - \hat{u_1}')^T||_F \\
&+ \|\frac{1}{n(n-1)}\sum_{c=2}^{K}\sum_{i\in C_k}(x_i - \hat{u_k})(x_i - \hat{u_k})^T\|_F \\
&\leq \frac{1}{n} + ||\frac{1}{n}\sum_{i\in C_1}[x_i x_i^T - \frac{1}{n_1}(\sum_{i\in C_1}x_i)x_i^T - \frac{1}{n_1}x_i(\sum_{i\in C_1}x_i)^T + \frac{1}{n_1^2}(\sum_{i\in C_1}x_i)(\sum_{i\in C_1}x_i)^T] - \\
&\frac{1}{n-1}\sum_{i\in C_1}[x_i' x_i'^T - \frac{1}{n_1-1}(\sum_{i\in C_1}x_i')x_i'^T - \frac{1}{n_1-1}x_i'(\sum_{i\in C_1}x_i')^T + \frac{1}{(n_1-1)^2}(\sum_{i\in C_1}x_i')(\sum_{i\in C_1}x_i')^T]||_F \\
&\leq ||\frac{1}{n}\sum_{i\in C_1}x_i x_i^T - \frac{1}{n-1}\sum_{i\in C_1}x_i' x_i'^T||_F + \frac{1}{n} \\
&+ ||\frac{2}{n}\frac{1}{n_1}(\sum_{i\in C_1}x_i)(\sum_{i\in C_1}x_i)^T - \frac{2}{n-1}\frac{1}{n_1-1}(\sum_{i\in C_1}x_i')(\sum_{i\in C_1}x_i')^T \\
&- \frac{1}{n}\frac{1}{n_1^2}n_1(\sum_{i\in C_1}x_i)(\sum_{i\in C_1}x_i)^T + \frac{1}{n-1}\frac{1}{(n_1-1)^2}(n_1-1)(\sum_{i\in C_1}x_i')(\sum_{i\in C_1}x_i')^T||_F \\
&\leq ||\frac{1}{n}x_j x_j^T - \frac{1}{n(n-1)}\sum_{i\in C_1, i\neq j}x_i x_i^T||_F \\
&\underbrace{+ ||\frac{1}{n}\frac{1}{n_1}(\sum_{i\in C_1}x_i)(\sum_{i\in C_1}x_i)^T - \frac{1}{n-1}\frac{1}{n_1-1}(\sum_{i\in C_1}x_i')(\sum_{i\in C_1}x_i')^T||_F}_{||\hat{B}(D)-\hat{B}(D')||_F} + \frac{1}{n} \\
&\leq \frac{1}{n} + \frac{1}{n(n-1)}(n-1) + \frac{5}{n} \leq \frac{7}{n}
\end{aligned}
$$

In total, for FDA, it satisfies Assumption 1 with $C_1 = 4$ and $C_2 = 7$.

Next, we will consider CCA. In CCA, there are two sets of variables $\{x_i\}_{i=1}^n \in \mathcal{R}^{d_1}$, and $\{y_i\}_{i=1}^n \in \mathcal{R}^{d_2}$. The $\mathcal{D}$ and $\mathcal{D}'$ are only different by deleting one data record (x_j, y_j). For convenience, we denote $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ and $\mathcal{D}' = \{x_i', y_i'\}_{i=1, i\neq j}^n$, note that $\{(x_i', y_i') = (x_i, y_i)\}$. Let $\hat{B_1} = \hat{\Sigma_x}, \hat{B_2} = \hat{\Sigma_y}$. For the matrix $\hat{A}(D)$, recall by the definition we have $||\hat{A}(D) - \hat{A}(D')||_F = 2\|\hat{\Sigma}_{xy} - \hat{\Sigma}'_{xy}\|_F$. Thus,

$$
\begin{aligned}
&\|\hat{\Sigma}_{xy} - \hat{\Sigma}'_{xy}\|_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i - \mu'_x)(y'_i - \mu'_y)^T||_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_i y_i^T - x_i\mu_y^T - \mu_x y_i^T + \mu_x\mu_y^T) - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i y_i^{'T} - x'_i\mu_y^{'T} - \mu'_x y_i^{'T} + \mu'_x\mu_y^{'T})||_F \\
&= ||\frac{1}{n}(\sum_{i=1}^{n} x_i y_i^T - n\mu_x\mu_y^T) - \frac{1}{n-1}(\sum_{i=1}^{n-1} x'_i y_i^{'T} - (n-1)\mu'_x\mu_y^{'T})||_F \\
&\leq ||\frac{1}{n}(\sum_{i\neq j} x_i y_i^T + x_j y_j^T) - \frac{1}{n-1}\sum_{i\neq j} x_i y_i^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j} x_i + x_j)(\sum_{i\neq j} y_i + y_j)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} y_i)^T||_F \\
&\leq ||\frac{1}{n}(\sum_{i\neq j} x_i y_i^T + x_j y_j^T) - \frac{1}{n-1}\sum_{i\neq j} x_i y_i^T||_F + ||\frac{1}{n^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} y_i)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} y_i)^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j} x_i)y_j^T||_F + ||\frac{1}{n^2}x_j(\sum_{i\neq j} y_i)^T||_F + ||\frac{1}{n^2}x_j y_j^T||_F \\
&\leq \frac{1}{n(n-1)}(n-1) + \frac{2n-1}{n^2(n-1)^2}(n-1)^2 + \frac{1}{n} + \frac{2}{n^2}(n-1) + \frac{1}{n^2} \leq \frac{6}{n}.
\end{aligned} \tag{30}
$$

For matrix $\hat{B}$, by the definition we have

$$
\|\hat{B}(D) - \hat{B}(D')\|_F \leq \|\hat{\Sigma}_x - \hat{\Sigma}'_x\|_F + \|\hat{\Sigma}_y - \hat{\Sigma}'_y\|_F.
$$

$$
\begin{aligned}
&\|\hat{\Sigma}_x - \hat{\Sigma}'_x\|_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i - \mu'_x)(x'_i - \mu'_x)^T||_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_i x_i^T - x_i\mu_x^T - \mu_x x_i^T + \mu_x\mu_x^T) - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i x_i^{'T} - x'_i\mu_x^{'T} - \mu'_x x_i^{'T} + \mu'_x\mu_x^{'T})||_F \\
&= ||\frac{1}{n}(\sum_{i=1}^{n} x_i x_i^T - n\mu_x\mu_x^T) - \frac{1}{n-1}(\sum_{i=1}^{n-1} x'_i x_i^{'T} - (n-1)\mu'_x\mu_x^{'T})||_F \\
&\leq ||\frac{1}{n}(\sum_{i\neq j} x_i x_i^T + x_j x_j^T) - \frac{1}{n-1}\sum_{i\neq j} x_i x_i^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j} x_i + x_j)(\sum_{i\neq j} x_i + x_j)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} x_i)^T||_F \\
&\leq ||\frac{1}{n}(\sum_{i\neq j} x_i x_i^T + x_j x_j^T) - \frac{1}{n-1}\sum_{i\neq j} x_i x_i^T||_F + ||\frac{1}{n^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} x_i)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j} x_i)(\sum_{i\neq j} x_i)^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j} x_i)x_j^T||_F + ||\frac{1}{n^2}x_j(\sum_{i\neq j} x_i)^T||_F + ||\frac{1}{n^2}x_j x_j^T||_F \\
&\leq \frac{1}{n(n-1)}(n-1) + \frac{2n-1}{n^2(n-1)^2}(n-1)^2 + \frac{1}{n} + \frac{2}{n^2}(n-1) + \frac{1}{n^2} \leq \frac{6}{n}
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
&\|\hat{\Sigma}_y - \hat{\Sigma}'_y\|_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_y)(y_i-\mu_y)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(y'_i-\mu'_y)(y'_i-\mu'_y)^T||_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(y_iy_i^T - y_i\mu_y^T - \mu_y y_i^T + \mu_y\mu_y^T) - \frac{1}{n-1}\sum_{i=1}^{n-1}(y'_iy_i'^T - y'_i\mu_y'^T - \mu'_y y_i'^T + \mu'_y\mu_y'^T)||_F \\
&= ||\frac{1}{n}(\sum_{i=1}^{n}y_iy_i^T - n\mu_y\mu_y^T) - \frac{1}{n-1}(\sum_{i=1}^{n-1}y'_iy_i'^T - (n-1)\mu'_y\mu_y'^T)||_F \\
&\le ||\frac{1}{n}(\sum_{i\neq j}y_iy_i^T + y_jy_j^T) - \frac{1}{n-1}\sum_{i\neq j}y_iy_i^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j}y_i + y_j)(\sum_{i\neq j}y_i + y_j)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j}y_i)(\sum_{i\neq j}y_i)^T||_F \\
&\le ||\frac{1}{n}(\sum_{i\neq j}y_iy_i^T + y_jy_j^T) - \frac{1}{n-1}\sum_{i\neq j}y_iy_i^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j}y_i)(\sum_{i\neq j}y_i)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j}y_i)(\sum_{i\neq j}y_i)^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j}y_i)y_j^T||_F + ||\frac{1}{n^2}y_j(\sum_{i\neq j}y_i)^T||_F + ||\frac{1}{n^2}y_jy_j^T||_F \\
&\le \frac{1}{n(n-1)}(n-1) + \frac{2n-1}{n^2(n-1)^2}(n-1)^2 + \frac{1}{n} + \frac{2}{n^2}(n-1) + \frac{1}{n^2} \\
&\le \frac{6}{n}
\end{aligned}
\tag{32}
$$

In total, for FDA, it satisfies Assumption 1 with $C_1 = 12$ and $C_2 = 12$.

In sliced inverse regression, we seek K vector $\{v_1, v_2, \cdots, v_K\}$ to represent the X. The $\mathcal{D}$ and $\mathcal{D}'$ are only different by deleting one data record (x_j, y_j) in the first class. For convenience, we denote $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ and $\mathcal{D}' = \{x'_i, y'_i\}_{i=1, i\neq j}^n$, note that $\{(x'_i, y'_i) = (x_i, y_i)\}$. Then, the data sensitivity of $\hat{A}$ and $\hat{B}$ are as follows.

$$
\begin{aligned}
&||\hat{A} - \hat{A}'||_F \\
&= |||[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(x_i-\mu_x)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i-\mu'_x)(x'_i-\mu'_x)^T] \\
&- [\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\hat{u_k})(x_i-\hat{u_k})^T - \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in C_k}(x'_i-\hat{u_k}')(x'_i-\hat{u_k}')^T]||_F \\
&\underbrace{\le ||\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(x_i-\mu_x)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(x'_i-\mu'_x)(x'_i-\mu'_x)^T||_F}_{||\hat{\Sigma}_x - \hat{\Sigma}'_x||_F} \\
&+ ||\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\hat{u_k})(x_i-\hat{u_k})^T - \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in C_k}(x'_i-\hat{u_k}')(x'_i-\hat{u_k}')^T||_F \\
&\le \frac{6}{n} + \frac{7}{n} \\
&\le \frac{13}{n}.
\end{aligned}
\tag{33}
$$

$$
\begin{aligned}
&||\hat{B}-\hat{B}'||_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(x_i-\mu_x)^T - \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i'-\mu_x')(x_i'-\mu_x')^T||_F \\
&= ||\frac{1}{n}\sum_{i=1}^{n}(x_ix_i^T - x_i\mu_x^T - \mu_x x_i^T + \mu_x\mu_x^T) - \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i'x_i'^T - x_i'\mu_x'^T - \mu_x'x_i'^T + \mu_x'\mu_x'^T)||_F \\
&= ||\frac{1}{n}(\sum_{i=1}^{n}x_ix_i^T - n\mu_x\mu_x^T) - \frac{1}{n-1}(\sum_{i=1}^{n-1}x_i'x_i'^T - (n-1)\mu_x'\mu_x'^T)||_F \\
&\le ||\frac{1}{n}(\sum_{i\neq j}x_ix_i^T + x_jx_j^T) - \frac{1}{n-1}\sum_{i\neq j}x_ix_i^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j}x_i + x_j)(\sum_{i\neq j}x_i + x_j)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j}x_i)(\sum_{i\neq j}x_i)^T||_F \\
&\le ||\frac{1}{n}(\sum_{i\neq j}x_ix_i^T + x_jx_j^T) - \frac{1}{n-1}\sum_{i\neq j}x_ix_i^T||_F + ||\frac{1}{n^2}(\sum_{i\neq j}x_i)(\sum_{i\neq j}x_i)^T - \frac{1}{(n-1)^2}(\sum_{i\neq j}x_i)(\sum_{i\neq j}x_i)^T||_F \\
&+ ||\frac{1}{n^2}(\sum_{i\neq j}x_i)x_j^T||_F + ||\frac{1}{n^2}x_j(\sum_{i\neq j}x_i)^T||_F + ||\frac{1}{n^2}x_jx_j^T||_F \\
&\le \frac{1}{n(n-1)}(n-1) + \frac{2n-1}{n^2(n-1)^2}(n-1)^2 + \frac{1}{n} + \frac{2}{n^2}(n-1) + \frac{1}{n^2} \le \frac{6}{n}.
\end{aligned} \tag{34}
$$

Thus, SIR satisfies Assumption 1 with $C_1 = 13$ and $C_2 = 6$.

In the following we will show that these models satisfy Assumption 2 when $\|x\|_2 \le 1$. Before that we first recall the following lemma for bounded random variable.

Lemma 4 (Bernstein Inequality [Vershynin, 2018])**.** Let $X_1, ..., X_N$ be independent random variables with $\|X_i\|_2 \le 1$, then for any $t > 0$ and for some absolute constant D_1 we have

$$
\mathbb{P}\Big(|\sum_{k=1}^{N}(X_k - \mathbb{E}X_k)| \ge t\Big) \le 2\exp\Big(-D_1\min\big\{\frac{t^2}{N}, t\big\}\Big). \tag{35}
$$

Thus, for PCA, by using Lemma 4 we have

$$
\|\hat{A} - A\|_2 \le \|\hat{A} - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\|_2 + \|\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T - \Sigma\|_2.
$$

For the second term, by Lemma 4 we have with probability at least $1-\zeta$

$$
\|\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T - \Sigma\|_{\infty,\infty} \le O(\frac{\sqrt{\log\frac{d}{\zeta}}}{\sqrt{n}} + \frac{\log d/\zeta}{n}).
$$

Thus we have

$$
\|\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T - \Sigma\|_2 \le \sqrt{d}\|\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T - \Sigma\|_{\infty,\infty} \le O(\frac{\sqrt{d\log\frac{d}{\zeta}}}{\sqrt{n}} + \frac{\sqrt{d}\log d/\zeta}{n}) \tag{36}
$$

For the first term we have

$$
\|\hat{A} - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\|_2 \le \|\hat{\mu}\hat{\mu}^T - \mu\mu^T\|_2 + 2\|\hat{\mu}-\mu\|_2 \le O(\frac{\sqrt{\log d/\zeta}}{\sqrt{n}}).
$$

This is due to that by corollary 7 in [Jin et al., 2019] and $\|x_i\|_2 \le 1$ we have with probability at least $1-\zeta$, $\|\hat{\mu}-\mu\|_2 \le O(\frac{\sqrt{\log d/\zeta}}{\sqrt{n}})$. In total we have $\|\hat{A} - A\|_2 \le O(\frac{\sqrt{\log d/\zeta}}{\sqrt{n}})$ with probability at least $1-\zeta$.

Next we will focus on CCA. For term A we have

$$\|\hat{A} - A\|_2 \leq O(\|\hat{\Sigma}_{xy} - \Sigma_{xy}\|_2).$$

By using a similar proof as in PCA case we have $\|\hat{A} - A\|_2 \leq O(\frac{\sqrt{d \log d/\zeta}}{\sqrt{n}})$ with probability at least $1 - \zeta$. The similar to term B. We omit here for simplicity.

For SIR, since the term B is just the covariance matrix, we have already show it in the PCA case. For term $\hat{A} = \hat{\Sigma}_x - E[\hat{\Sigma}_{(x|y)}]$. Thus

$$\|\hat{A} - A\|_2 \leq \|\hat{\Sigma}_x - \Sigma_x\|_2 + \|E[\hat{\Sigma}_{(x|y)}] - E[\Sigma_{(x|y)}]\|_2.$$

The first term is the same as the PCA case. So $\|\hat{\Sigma}_x - \Sigma_x\|_2 \leq \frac{\sqrt{d \log d/\zeta}}{\sqrt{n}}$. For the second term by the form of $E[\hat{\Sigma}_{(x|y)}]$ and the previous results we can see that

$$\|E[\hat{\Sigma}_{(x|y)}] - E[\Sigma_{(x|y)}]\|_2 \leq O(\sum_{k=1}^{K} \frac{\sqrt{dn_k \log d/\zeta}}{n}) \leq O(\sum_{k=1}^{K} \frac{K\sqrt{d \log d/\zeta}}{\sqrt{n}}). \tag{37}$$

Thus we finish the proof.

For the term $\rho(E_A, k)$ and $\rho(E_B, k)$. Note that since x is $O(1)$-subGaussian. Thus, by Lemma 12 in [Gao et al., 2015] we can show the statements for PCA and CCA. For SIR, as we mentioned previously, it is just a combination of the in-class covariance matrix estimation. Since it is true for PCA, thus we can also see it is true for SIR. □

4 Proof of CDP

For both Algorithm 1, 2 and 3 we can see that by Assumption 1 and Gaussian mechanism in each iteration calculating $\tilde{A}^t$ and $\tilde{B}^t$ is $\frac{\rho}{m}$-CDP. Thus by the Composition Theorem we can see the whole algorithm is ρ-CDP.

Note that since the high dimensional sparse case is more general. Thus, we first show the proofs of the Theorems for the high dimensional sparse case.

5 Proof of Theorem 7

Instead of proof Theorem 7, we propose the following assumption and show the following result.

Assumption 3. For sufficient large n, there are constants $c \geq 0$ and $0 \leq b < \min_{j \in [d]} \frac{\lambda_j(F)}{2\lambda_j^2(F)+1}$ such that $\frac{\tilde{\iota}(k')}{\text{cr}(k')} \leq \frac{b}{2}$ and $\rho(E_B, k') \leq \frac{c}{2}\lambda_{\min}(B)$, where $\tilde{\iota}(k')$, $\text{cr}(k')$ are defined as

$$\begin{aligned} \text{cr}(k') &= \inf_{F:|F| \leq k'} \text{cr}(A_F, B_F), \\ \tilde{\iota}(k') &= \rho(E_A, k') + \rho(E_B, k'). \end{aligned} \tag{38}$$

Theorem 13. Under Assumption 3, given any failure probability $\zeta > 0$, if n is sufficiently large such that, $n \geq \Omega(\max\{\frac{\sqrt{k'm \log d \log \frac{m}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{k'm \log d \log \frac{m}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$.

Then with probability at least $1 - \zeta$, there exists constants a and c such that for all $t \in [m]$,

$$(1-a)\lambda_j(F) \leq \tilde{\lambda}_j^t(F) \leq (1+a)\lambda_j(F), \tag{39}$$

$$(1-c)\lambda_j(B_F) \leq \lambda_j(\tilde{B}_F^t) \leq (1+c)\lambda_j(B_F), \tag{40}$$

$$C_{\text{lower}}\kappa(B) \leq \kappa(\tilde{B}_F^t) \leq C_{\text{upper}}\kappa(B) \tag{41}$$

where $C_{\text{lower}} = \frac{1-c}{1+c}$, $C_{\text{upper}} = \frac{1+c}{1-c}$, c is the constant as in Assumption 3. Furthermore, we have

$$\tilde{\lambda}_2^t(F) \leq \gamma\tilde{\lambda}_1^t(F), \tag{42}$$

where $\gamma = \frac{(1+a)\lambda_2(F)}{(1-a)\lambda_1(F)}$.

Based on Assumption 2 we can show that when $n \geq \Omega(\max\{\frac{k'}{b^2\text{cr}^2(k')}, \frac{k'}{c^2\lambda^2_{\min}(B)}\})$ then it satisfies Assumption 3. Then by Theorem 13 we can **proof Theorem 7**.

5.1 Proof of Theorem 13

Before showing the proof of Theorem 13, we first prove the following key result.

Theorem 14. Under Assumption 3, given any iteration t and failure probability ζ, if n is sufficiently large such that, $n \geq \tilde{\Omega}(\max\{\frac{\sqrt{k'm\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{k'm\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$. Then with probability at least $1-\zeta$, there are constants $b, c > 0$ such that

$$\frac{\iota(k')}{\text{cr}(k')} \leq b \tag{43}$$

$$\rho(E_B + Z_2^t, k') \leq c\lambda_{\min}(B), \tag{44}$$

where $\iota(k')$ is defined as following

$$\iota(k') = \sqrt{\rho(E_A + Z_1^t, k')^2 + \rho(E_B + Z_2^t, k')^2}.$$

Next, we recall the following two lemmas.

Lemma 5. [[Stewart, 1979]] Let J and $J + E_J$ be $d \times d$ symmetric matrices. Then, for all $k \in \{1, \cdots, d\}$, we have

$$\lambda_k(J) + \lambda_{\min}(E_J) \leq \lambda_k(J + E_J) \leq \lambda_k(J) + \lambda_{\max}(E_J).$$

Lemma 6. [[Stewart, 1979]] Let (J, K) be a symmetric definite matrix pair with generalized eigenvalues $\lambda_1(J, K) \geq \lambda_2(J, K) \geq \cdots \geq \lambda_d(J, K)$. Let $(J + E_J, K + E_K)$ be the perturbed matrix pair. If E_J and E_K satisfy

$$\epsilon = \sqrt{||E_J||_2^2 + \|E_K\|_2^2} \leq \text{cr}(J, K),$$

where cr is defined by Assumption 2. Then, we have $\lambda_1(J + E_J, K + E_K) \geq \cdots \geq \lambda_d(J + E_J, K + E_K)$. And, the following inequality holds,

$$\frac{\lambda_k(J,K)\text{cr}(J,K) - \epsilon}{\text{cr}(J,K) + \epsilon\lambda_k(J,K)} \leq \lambda_k(J + E_J, K + E_K) \leq \frac{\lambda_k(J,K)\text{cr}(J,K) + \epsilon}{\text{cr}(J,K) - \epsilon\lambda_k(J,K)}.$$

Proof of Theorem 13. By Theorem 14 we can see that when n is sufficiently large, Theorem 14 holds for all $t \in [m]$ with probability at least $1 - \zeta$.

We fix an iteration t and for convenience we omit the superscript t. Then, in Lemma 6 we take $J = A_F$, $K = B_F$, $E_J = E_A + Z_1$ and $E_K = E_B + Z_2$. Following $\frac{\epsilon(k')}{\text{cr}(k')} \leq b$ in Theorem 14 and Lemma 6 we have

$$\frac{\lambda_k(F) - b}{1 + b\lambda_k(F)} \leq \lambda_k(\tilde{F}) \leq \frac{\lambda_k(F) + b}{1 - b\lambda_k(F)}.$$

Define a constant a satisfying

$$1 > a \geq \frac{b/\lambda_k(F) + b\lambda_k(F)}{1 - b\lambda_k(F)},$$

then we have $(1-a)\lambda_k(F) \leq \tilde{\lambda}_k(F) \leq (1+a)\lambda_k(F)$. Note that since $b < \frac{\lambda_k(F)}{2\lambda_k^2(F)+1}$, there always exists such an a. Thus, we derive the inequality (39).

We can get (40) via Lemma 5 with $J = B$ and $E_J = E_B + Z_2$.

By (40) we have

$$(1-c)\lambda_{\max}(B_F) \leq \lambda_{\max}(\tilde{B}_F) \leq (1+c)\lambda_{\max}(B_F),$$

and

$$(1-c)\lambda_{\min}(B_F) \leq \lambda_{\min}(\tilde{B}_F) \leq (1+c)\lambda_{\min}(B_F),$$

also $\lambda_{\min}(B_F) \leq \lambda_{\max}(B_F)$. Thus, we have

$$(\frac{1-c}{1+c})\frac{\lambda_{\max}(B_F)}{\lambda_{\min}(B_F)} \leq \frac{\lambda_{\max}(\tilde{B}_F)}{\lambda_{\min}(\tilde{B}_F)} \leq (\frac{1+c}{1-c})\frac{\lambda_{\max}(B_F)}{\lambda_{\min}(B_F)}.$$

Also by the definition we know that $\lambda_{\max}(B_F) \leq \lambda_{\max}(B)$ and $\lambda_{\min}(B_F) \geq \lambda_{\min}(B)$. Thus, the inequality (41) holds. Also, following the inequality (39), we have

$$(\frac{1-a}{1+a})\frac{\lambda_1(F)}{\lambda_2(F)} \leq \frac{\tilde{\lambda}_1(F)}{\tilde{\lambda}_2(F)}.$$

Thus, the inequality (42) holds. □

Proof of Theorem 14. By the definition and Assumption 3 it is easy to see that it is sufficient to show that

$$\rho(Z_1^t, k') + \rho(Z_2^t, k') \leq \text{cr}(k')\frac{b}{2}, \tag{45}$$

$$\rho(Z_2^t, k') \leq \frac{c\lambda_{\min}(B)}{2}. \tag{46}$$

To show these, we will use the following lemma.

Lemma 7. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric Gaussian random matrix where each entry is sampled from $\mathcal{N}(0, \sigma^2)$. Then we have with probability at least $1-\zeta$,

$$\sup_{\|u\|_2=1, \|u\|_0\leq s} |u^T A u| \leq O(\sigma\sqrt{s\log\frac{d}{\zeta}}). \tag{47}$$

By Lemma 7 we have that when $n \geq \Omega(\frac{\sqrt{k'm\log d\log\frac{1}{\zeta}}}{\text{cr}(k')b\sqrt{\rho}})$ and $n \geq \Omega(\frac{\sqrt{k'm\log d\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}})$, then with probability at least $1-\zeta$, (45) and (46) hold.

□

Proof of Lemma 7. Let $\mathcal{S}_\epsilon$ be an ϵ-net of S. By Lemma 3 we have $\mathcal{N}(S,\epsilon) \leq e^K \cdot \frac{cd}{s\epsilon}$.

For a fixed $u \in S$, there is a $u_0 \in \mathcal{S}_\epsilon$ such that $\|u-u_0\|_2 \leq \epsilon$. Thus, we have with probability at least $1-\frac{\zeta}{2}$,

$$\begin{aligned}
|u^T A u| &= |u^T A(u-u_0) + u^T A u_0| \\
&= |u^T A(u-u_0) + (u-u_0)^T A u_0 + u_0^T A u_0| \\
&\leq u_0^T A u_0 + 2\epsilon\|A\|_2 \\
&\leq |u_0^T A u_0| + 2\epsilon C_1\sqrt{d}\sqrt{\log\frac{1}{\zeta}}\sigma,
\end{aligned}$$

where C_1 is a universal constant and the last inequality is due to Theorem 4.4.5 in [Vershynin, 2018]. Thus we have

$$\sup_{\|u\|_2=1, \|u\|_0\leq s} |u^T A u| \leq \sup_{u_0\in\mathcal{S}_\epsilon} |u_0^T A u_0| + 2\epsilon\|A\|_2$$

Next we will bound the term $\sup_{u_0\in\mathcal{S}_\epsilon} |u_0^T A u_0|$. For a fixed $u_0 \in \mathcal{S}_\epsilon$,

$$u_0^T A u_0 = \sum_{i=1,j=1,i\geq j}^{d} A_{ij}u_{0i}u_{0j} + \sum_{i=1.j=1,i<j}^{d} A_{ij}u_{0i}u_{0j}.$$

$\sum_{i=1,j=1,i\geq j}^{d} A_{ij}u_{0i}u_{0j}$ and $\sum_{i=1.j=1,i<j}^{d} A_{ij}u_{0i}u_{0j}$ are all sums of independent Gaussian random variables. Thus, by the tails bound of Gaussian distribution we have

$$\mathbb{P}(|\sum_{i=1,j=1,i\geq j}^{d} A_{ij}u_{0i}u_{0j} \geq t|) \leq 2\exp(-\frac{t^2}{2\sigma^2}).$$

Thus, we have

$$\mathbb{P}(\sup_{u_0\in\mathcal{S}_\epsilon} |u_0^T A u_0| \geq t) \leq \sum_{u_0\in\mathcal{S}_\epsilon} \mathbb{P}(|u_0^T A u_0| \geq t) \leq 2e^s \frac{cd}{s\epsilon}\exp(-\frac{t^2}{2\sigma^2}).$$

That is with probability at least $1-\frac{\zeta}{2}$, $\sup_{u_0\in\mathcal{S}_\epsilon} |u_0^T A u_0| \leq C_2\sigma\sqrt{s\log\frac{d}{s\epsilon\zeta}}$.

For the operator norm of a symmetric Gaussian matrix we have the following lemma:

Lemma 8 ([Biswas et al., 2020])**.** Let $A\in\mathbb{R}^{d\times d}$ be a symmetric Gaussian noise with each entry $A_{ij}\sim\mathcal{N}(0,\sigma^2)$. Then with probability at least $1-\beta$, $\|A\|_2 \leq O(\sigma\sqrt{d\log\frac{1}{\beta}})$.

In total, we have with probability at least $1-\zeta$,

$$|u^T A u| \leq O(\epsilon\sigma\sqrt{d\log\frac{1}{\zeta}} + \sigma\sqrt{s\log\frac{d}{s\epsilon\zeta}}).$$

Take $\epsilon = \frac{\sqrt{s}}{\sqrt{d}}$ we can get the result. □

6 Proof of Theorem 8

Theorem 15 (Full version of Theorem 8)**.** Under Theorem 7 with $k' = 2k+s$ and choose $k = Cs$ for sufficiently large $C>0$. In addition, choose the stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1+2\sqrt{\frac{s}{k}}+2\frac{s}{k}}\sqrt{1-\frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B)+\gamma}} < \frac{1}{2}.$$

Assume that n is sufficiently large such that

$$n \geq \tilde{\Omega}(\max\{\frac{k'}{\text{cr}(k')^2\lambda_{gap}^2}, \frac{k'}{(\max\{\theta(A,B),\sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(k'))^2}\frac{\sqrt{s\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{s\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}, \frac{\sqrt{s\log\frac{1}{\zeta}}}{\text{cr}(k')\lambda_{gap}\sqrt{\rho}}, \frac{\sqrt{s\log\frac{1}{\zeta}}}{\lambda_{gap}\text{cr}(k')\theta(A,B)\sqrt{\rho}}\})$$

with λ_{gap} in (12) for any given failure probability $\zeta>0$. Then in Algorithm 3, if we set $m=O(\log n)$ and if the input k-sparse vector v_0 with $\|v_0\|_2$ satisfying $\frac{|\langle v_s^*, v_0\rangle|}{\|v_s^*\|_2} \geq 1-\frac{\theta(A,B)}{2}$ with $\theta(A,B)$ given in (10). We have the following with probability at least $1-\zeta$,

$$1-\frac{\langle v_s^*, v_m\rangle}{\|v_s^*\|_2} \leq O(\frac{\theta(A,B)}{(1-\nu)^2}(\frac{\tilde{\iota}^2(k')}{\lambda_{gap}^2\text{cr}^2(k')} + \frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log n\log d\log\frac{1}{\zeta}}{n^2\rho})). \tag{48}$$

Similarly, for the deterministic setting where $E_A = E_B = 0$ and $\hat{A}=A, \hat{B}=B$, if n is sufficiently large and with some additional mild assumptions. If we set some appropriate parameters in Algorithm 3, with probability at least $1-\zeta$

$$1-\frac{\langle \tilde{v}_s, v_t\rangle}{\|\tilde{v}_s\|_2} \leq \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d\log\frac{1}{\zeta}}{n^2\rho}). \tag{49}$$

In stead of proof Theorem 15 we proof the following theorem. Note that by Theorem 16 and Assumption 2 we can get Theorem 15 .

Theorem 16. Under Theorem 7 with $k' = 2k + s$ and choose $k = Cs$ for sufficiently large $C > 0$. In addition, choose the stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 + 2\sqrt{\frac{s}{k}} + 2\frac{s}{k}}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B) + \gamma}} < \frac{1}{2}.$$

. Assume that $\lambda_{gap} > \frac{\iota(\tilde{k}')}{\text{cr}(k')}$, $\tilde{\iota}(k') \leq O(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(k'))$ with λ_{gap} in (12) and n is sufficiently large such that

$$n \geq \tilde{\Omega}(\max\{\frac{\sqrt{s\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{s\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}, \frac{\sqrt{s\log\frac{1}{\zeta}}}{\text{cr}(k')\lambda_{gap}\sqrt{\rho}}, \frac{\sqrt{s\log\log\frac{1}{\zeta}}}{\lambda_{gap}\text{cr}(k')\theta(A,B)\sqrt{\rho}}\})$$

for any given failure probability $\zeta > 0$. Then in Algorithm 3, if we set $T = O(\log n)$ and if the input k-sparse vector v_0 with $\|v_0\|_2$ satisfying $\frac{|\langle v_s^*, v_0\rangle|}{\|v_s^*\|_2} \geq 1 - \frac{\theta(A,B)}{2}$ with $\theta(A,B)$ given in (10). We have the following with probability at least $1 - \zeta$,

$$1 - \frac{\langle v_s^*, v_m\rangle}{\|v_s^*\|_2} \leq O(\frac{\theta(A,B)}{(1-\nu)^2}(\frac{\tilde{\iota}^2(k')}{\lambda_{gap}^2\text{cr}^2(k')} + \frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log n\log d\log\frac{1}{\zeta}}{n^2\rho})). \tag{50}$$

Similarly, for the deterministic setting where $E_A = E_B = 0$ and $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and with some additional mild assumptions. If we set some appropriate parameters in Algorithm 3, with probability at least $1 - \zeta$

$$1 - \frac{\langle \tilde{v}_s, v_t\rangle}{\|\tilde{v}_s\|_2} \leq \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d\log\frac{1}{\zeta}}{n^2\rho}). \tag{51}$$

Proof of Theorem 16. We mainly focus on the stochastic setting. The result for the deterministic setting is just a special case where $\tilde{\iota}(k') = 0$ since $E_A = E_B = 0$.

In each iteration t we denote the vector $v^t(F)$ as the solution of the following GEP restricted to a superset of V.

$$v^t(F) = \arg\max_{v\in\mathbb{R}^d} v^T\tilde{A}^t v, \text{ s.t. } v^T\tilde{B}^t v = 1, \text{supp}(v) \subseteq F. \tag{52}$$

We also denote $y^* = \frac{v_s^*}{\|v_s^*\|_2}$ and $y^t(F) = \frac{v^t(F)}{\|v^t(F)\|_2}$. The following theorem shows the error of the vector in each iteration.

Theorem 17. Under Theorem 13 with $k' = 2k + s$ and choose $k = Cs$ for sufficiently large $C > 0$. In addition, choose the stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 + 2\sqrt{\frac{s}{k}} + 2\frac{s}{k}}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B) + \gamma}} < \frac{1}{2}.$$

Suppose in the t-th iteration, v_{t-1} satisfies $\langle y^t(F), v_{t-1}\rangle \geq 1 - \theta(A,B)$, where $\theta(A,B)$ is in (10). Then we have

$$\sqrt{1 - |\langle y^*, v_t\rangle|} \leq \nu\sqrt{1 - |\langle y^*, v_{t-1}\rangle|} + \sqrt{20}\sqrt{1 - |\langle y^t(F), y^*\rangle|}. \tag{53}$$

To proof Theorem 16 we will show the following lemmas:

Lemma 9. Under Theorem 17, if $\tilde{\iota}(k') \leq O(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(k'))$ and $n \geq \Omega(\frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{\min\{\theta(A,B),\sqrt{\theta(A,B)}\}\sqrt{\rho}})$, then for all $t \in [m]$, if $|\langle v^*, v_{t-1}\rangle|/\|v^*\|_2 \geq 1 - \frac{\theta(A,B)}{2}$, then

$$\langle y^t(F), v_{t-1}\rangle \geq 1 - \theta(A,B). \tag{54}$$

Lemma 10. Under Theorem 17 and Lemma 9, for all $t \in [m]$, if $\langle y^t(F), v_{t-1}\rangle \geq 1 - \theta(A,B)$, then

$$\sqrt{1-|\langle y^*, v_t\rangle|} \leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + O(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \tag{55}$$

and

$$\frac{|\langle v_s^*, v_t\rangle|}{\|v_s^*\|_2} \geq 1 - \frac{\theta(A,B)}{2}. \tag{56}$$

Combining these two lemmas we have under the assumptions, for each $t \in [m]$,

$$\sqrt{1-\frac{|\langle v_s^*, v_m\rangle|}{\|v_s^*\|_2}} \leq \nu^m\sqrt{\theta(A,B)} + O(\frac{1}{1-\nu}(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}})). \tag{57}$$

Take $m = O(\log n)$ we finish the proof.

Proof of Lemma 9. Since

$$\langle y^t(F), v_{t-1}\rangle \geq \frac{|\langle v^*, v_{t-1}\rangle|}{\|v^*\|_2} - \|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2,$$

it is sufficiently to show that $\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 \leq \frac{\theta(A,B)}{2}$.

To show this, we will provide a stronger statement.

Lemma 11. Under Theorem 13, for each $t \in [m]$ if $n \geq \Omega(\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{\text{cr}(k')\sqrt{\rho}})$,

$$\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 \leq O(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}).$$

Proof of Lemma 11. By Theorem 4.3 in [Stewart, 1979] we have

$$\frac{\|v^t(F) - v_s^*\|_2}{\|v_s^*\|_2} \leq \frac{\iota(k')}{\Delta\tilde{\lambda}^t\text{cr}(\tilde{A}_F^t, \tilde{B}_F^t)},$$

where $\Delta\tilde{\lambda}^t = \min_{k>1}\chi\{\lambda_1(F), \tilde{\lambda}_k^t(F)\} > 0$ and

$$\chi\{\lambda_1(F), \tilde{\lambda}_k^t(F)\} = \frac{|\lambda_1(F) - \tilde{\lambda}_k^t(F)|}{\sqrt{1+\lambda_1(F)^2}\sqrt{1+\tilde{\lambda}_k^t(F)^2}}.$$

Thus, we have

$$\begin{aligned}\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 &= \frac{1}{\|v^t(F)\|_2\|v_s^*\|_2}\|v^t(F)\|v_s^*\|_2 - v_s^*\|v(F)\|_2\|_2 \\ &\leq \frac{2}{\|v_s^*\|_2}\|v^t(F) - v_s^*\|_2 \\ &\leq \frac{2}{\|v_s^*\|_2}\|v^t(F) - v_s^*\|_2 \leq 2\frac{\iota(k')}{\Delta\tilde{\lambda}^t\text{cr}(\tilde{A}_F^t, \tilde{B}_F^t)}.\end{aligned}$$

By Theorem 13 we can see that $\Delta\tilde{\lambda} \geq \lambda_{gap}$, and by Theorem 2.4 in [Stewart, 1979] we have $\text{cr}(\tilde{A}_F^t, \tilde{B}_F^t) \geq \text{cr}(k') - \iota(k')$. Thus we have

$$\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 \leq O(\frac{\iota(k')}{\lambda_{gap}(\text{cr}(k') - \iota(k'))}).$$

By the definition of $\iota(k')$ we have $\iota(k') \leq \tilde{\iota}(k') + O(\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}})$. Since $\frac{\tilde{\iota}(k')}{\text{cr}(k')} \leq O(1)$ and $n \geq \Omega(\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{\text{cr}(k')\sqrt{\rho}})$, we have

$$\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 \leq O(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}).$$

□

Thus, we can see that when $\tilde{\iota}(k') \leq O(\theta(A,B)\lambda_{gap}\text{cr}(k'))$ and $n \geq \Omega(\frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\delta}\log\frac{m}{\zeta}}}{\theta(A,B)\epsilon})$ then $\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v_s^*}{\|v_s^*\|_2}\|_2 \leq \frac{\theta(A,B)}{2}$. □

Proof of Lemma 10 . By Theorem 17 and Lemma 11 we have

$$\begin{aligned}
\sqrt{1-|\langle y^*, v_t\rangle|} &\leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + \sqrt{20}\sqrt{1-|\langle y^t(F), y^*\rangle|} \\
&\leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + O(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \\
&\leq \nu\sqrt{\theta(A,B)} + O(\frac{\tilde{\iota}(k')}{\lambda_{gap}\text{cr}(k')} + \frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \\
&\leq \frac{\sqrt{\theta(A,B)}}{\sqrt{2}},
\end{aligned}$$

where the last inequality is due to that $\nu < \frac{1}{2}$, $\tilde{\iota}(k') \leq O(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(k'))$ and $n \geq \tilde{\Omega}(\frac{1}{\lambda_{gap}\text{cr}(k')}\frac{\sqrt{sm\log\frac{m}{\zeta}}}{\min\{\theta(A,B),\sqrt{\theta(A,B)}\}\sqrt{\rho}})$. □

□

6.1 Proof of Theorem 17

Before providing the proof we first show the following lemma.

Lemma 12. Assume that Theorem 13 holds, consider in the t-th iteration, let $F \subset \{1, 2, \cdots, d\}$, and $V \subset F$, $|F| = k'$. Given any $\tilde{v}$, such that $||\tilde{v}||_2 = 1$ and $\langle\tilde{v}, y(F)\rangle > 0$. Let $\rho = \frac{\tilde{v}^T\tilde{A}^t\tilde{v}}{\tilde{v}^T\tilde{B}^t\tilde{v}} = \frac{\tilde{v}^T\tilde{A}_F^t\tilde{v}}{\tilde{v}^T\tilde{B}_F^t\tilde{v}}$, and let $v' = \frac{C^t\tilde{v}}{||C^t\tilde{v}||_2} = \frac{C_F^t\tilde{v}}{||C_F^t\tilde{v}||_2}$, where

$$C^t = I + \frac{\eta}{\rho}(\tilde{A}^t - \rho\tilde{B}^t),$$

and $\eta > 0$ is a positive constant. Define η such that

$$\eta\lambda_{\max}(B) < \frac{1}{1+c}.$$

Denote $\delta = 1 - \langle y^t(F), \tilde{v}\rangle$, and $1 - \delta \geq 1 - \theta(A,B)$, where

$$\theta(A,B) = \min\{\frac{1}{8C_{\text{upper}}\kappa(B)}, \frac{1/\gamma - 1}{3C_{\text{upper}}\kappa(B)}, \frac{1-\gamma}{30(1+c)C_{\text{upper}}^2\eta\lambda_{\max}(B)\kappa^2(B)\{C_{upper}\kappa(B)+\gamma\}}\}.$$

Then, under the conditions in Theorem 13 we have

$$\langle v', y^t(F)\rangle \geq \langle\tilde{v}, y^t(F)\rangle + \frac{1+c}{8}\eta\lambda_{\min}(B)\{1 - \langle\tilde{v}, y^t(F)\rangle\}\frac{1-\gamma}{C_{\text{upper}}\kappa+\gamma}. \tag{58}$$

We also need the following lemma.

Lemma 13 ([Tan et al., 2018])**.** Consider y' with $F' = \text{supp}(y')$ and $|F'| = \bar{k}$. Consider y and let $F = \text{supp}(y, k)$ be the indices of y with the largest k absolute values with $|F|$ = k. If $||y'||_2 = ||y||_2 = 1$, then

$$|\text{truncate}(y,F)^T y'| \geq |y^T y'| - \sqrt{\frac{\bar{k}}{k}}\min\{\sqrt{1-(y^T y')^2}, [1+\sqrt{\frac{\bar{k}}{k}}][1-(y^T y')^2]\}$$

Proof of Theorem 17. Recall that, in Algorithm 3, we truncate the v_t' and define it to $\hat{v}_t$. Also, we defined that $v_t = \frac{\hat{v}_t}{||\hat{v}_t||_2}$, and $||v_t'||_2 = 1$. Since the $\hat{v}_t$ is the truncated version of v_t', we have $||\hat{v}_t||_2 \leq 1$, and thus $|\langle y^*, v_t\rangle| \geq |\langle y^*, \hat{v}_t\rangle|$. We then evaluate the Algorithm 3 in each iteration t.

Assume that $k > s$, where s is the cardinality of the support of $y^* = \frac{v_s^*}{||v_s^*||_2^2}$, k is the truncation parameter in Algorithm 3. Let $k' = 2k + s$. Let $F_{t-1} = \text{supp}(v_{t-1})$, and $F_t = \text{supp}(v_t)$. And let $F = F_{t-1} \cup F_t \cup V$. Recall that $y^* = \frac{v*}{||v^*||_2}$. We have $|F| \leq k' = 2k + s$, since $|F_t| = |F_{t-1}| = k$. Let

$$v_t' = \frac{C_F^t v_{t-1}}{||C_F^t v_{t-1}||_2},$$

where the C_F^t is the submatrix of C_F^t indexed by F, and then the v_t' is equivalent to the value of that in Algorithm 3. Now, we prove the conclusion.

Since we assume Theorem 13 holds, following the Lemma 12 with F, $\tilde{v} = v_{t-1}$, and $v' = v_t'$, we obtain that

$$\langle y^t(F), v_t' \rangle \geq \langle y^t(F), v_{t-1} \rangle + \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\{1 - \langle y^t(F), v_{t-1}\rangle\}\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}.$$

Thus, we have

$$1 - \langle y^t(F), v_t' \rangle \leq \{1 - \langle y^t(F), v_{t-1}\rangle\}\{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}\}.$$

By subtracting that $||y(F)||_2 = 1$ and $||v_t'||_2 = 1$, we have

$$||y^t(F) - v_t'||_2 \leq ||y^t(F) - v_{t-1}||_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}}.$$

Then, we calculate the difference to the optimal solution, we have

$$\begin{aligned}
||y^* - v_t'||_2 &\leq ||y^t(F) - v_t'||_2 + ||y^t(F) - y^*||_2 \\
&\leq ||y^t(F) - v_{t-1}||_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + ||y^t(F) - y^*||_2 \\
&\leq ||y^* - v_{t-1}||_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + 2||y^t(F) - y^*||_2.
\end{aligned}$$

Thus, we have

$$\sqrt{1 - |\langle y^*, v_t' \rangle|} \leq \sqrt{1 - |\langle y^*, v_{t-1}\rangle|}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + 2\sqrt{1 - |\langle y^t(F), y^*\rangle|}.$$

We define

$$\nu = \sqrt{1 + 2[\sqrt{\frac{s}{k}} + \frac{s}{k}]}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\text{min}}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}}.$$

By using Lemma 13 and assume $k > s$, we have

$$\begin{aligned}
\sqrt{1 - |\langle y^*, \hat{v}_t \rangle|} &\leq \sqrt{1 - |\langle y^*, v_t' \rangle| + [\sqrt{\frac{s}{k}} + \frac{s}{k}](1 - |\langle y^*, v_t' \rangle|)^2} \\
&\leq \sqrt{1 - |\langle y^*, v_t' \rangle|}\sqrt{1 + [\sqrt{\frac{s}{k}} + \frac{s}{k}](1 + |\langle y^*, v_t' \rangle|)} \\
&\leq \sqrt{1 - |\langle y^*, v_t' \rangle|}\sqrt{1 + 2[\sqrt{\frac{s}{k}} + \frac{s}{k}]} \\
&\leq \nu\sqrt{1 - |\langle y^*, v_{t-1} \rangle|} + \sqrt{20}\sqrt{1 - |\langle y^t(F), y^* \rangle|}.
\end{aligned}$$

Thus, we can get

$$\begin{aligned}
\sqrt{1 - |\langle y^*, v_t \rangle|} &\leq \sqrt{1 - |\langle y^*, \hat{v}_t \rangle|} \\
&\leq \nu\sqrt{1 - |\langle y^*, v_{t-1} \rangle|} + \sqrt{20}\sqrt{1 - |\langle y^t(F), y^* \rangle|}.
\end{aligned}$$

□

Proof of Lemma 12. For convenience we omit the superscript t in the proof.

The proof follow [Tan et al., 2018]. We intend to derive the lower bound of $y(F)^T v'$. Follow the definition, we have

$$\langle y(F), v' \rangle = \underbrace{y(F)^T C_F \tilde{v}}_{M} \cdot \underbrace{||C_F \tilde{v}||_2^{-1}}_{N}. \tag{59}$$

We first clarify some important conclusions to prove the bound conveniently. Recall that $F \subseteq \{1, \cdots, d\}$ is some set with cardinality $|F| = k'$, $y(F)$ is proportional to the largest generalized eigenvector of $(\tilde{A}_F, \tilde{B}_F)$. In this proof, we denote $\tilde{\kappa}$ to represent $\kappa(\tilde{B}_F)$. Besides, we use $||v||_{\tilde{B}_F}^2$ to indicate $v^T \tilde{B}_F v$.

We define a basis vector of the space spanned by $\tilde{B}_F$ called $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_{k'}$. Let $\boldsymbol{\xi}_j$ be the jth generalized eigenvector of $(\tilde{A}_F, \tilde{B}_F)$ corresponding to $\tilde{\lambda}_j(F)$ such that (note that such $\{\xi\}_i$ exists [Golub and Van Loan, 1996])

$$\boldsymbol{\xi}_j^T \tilde{B}_F \boldsymbol{\xi}_k = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

Assume that $\tilde{v} = \sum_{j=1}^{k'} \alpha_j \boldsymbol{\xi}_j$ and we have $y(F) = \frac{\boldsymbol{\xi}_1}{||\boldsymbol{\xi}_1||_2}$. By assumption, we have $y(F)^T \tilde{v} = 1 - \delta$. This implies that $||y(F) - \tilde{v}||_2^2 = 2\delta$. Also, note that

$$\begin{aligned} ||\tilde{v} - y(F)||_{\tilde{B}_F}^2 &= ||\tilde{v} - \alpha_1 \boldsymbol{\xi}_1 - \{y(F) - \alpha_1 \boldsymbol{\xi}_1\}||_{\tilde{B}_F}^2 \\ &= ||\tilde{v} - \alpha_1 \boldsymbol{\xi}_1||_{\tilde{B}_F}^2 + ||y(F) - \alpha_1 \boldsymbol{\xi}_1||_{\tilde{B}_F}^2 - 2(y(F) - \alpha_1 \boldsymbol{\xi}_1)^T \tilde{B}_F (\tilde{v} - \alpha_1 \boldsymbol{\xi}_1). \end{aligned} \tag{60}$$

Since $y(F) - \alpha_1 \boldsymbol{\xi}_1$ is orthogonal to $\tilde{v} - \alpha_1 \boldsymbol{\xi}_1$ under the normalization of $\tilde{B}_F$ (by the definition of $y(F)$ and $\tilde{v}$), we have

$$\sum_{j=2}^{k'} \alpha_j^2 = ||\tilde{v} - \alpha_1 \boldsymbol{\xi}_1||_{\tilde{B}_F}^2 \leq ||\tilde{v} - y(F)||_{\tilde{B}_F}^2 \leq 2\lambda_{\max}(\tilde{B}_F)\delta, \tag{61}$$

verified by the fact that $||y(F) - \tilde{v}||_2^2 = 2\delta$. Similarly, we have

$$\begin{aligned} &\sum_{j=1}^{k'} \alpha_j^2 = ||\tilde{v}||_{\tilde{B}_F}^2 \geq \lambda_{\min}(\tilde{B}_F) = \lambda_{\max}(\tilde{B}_F)/\tilde{\kappa}, \\ &\alpha_1^2 \geq \lambda_{\max}/\tilde{\kappa} - \sum_{j=1}^{k'} \alpha_j^2 \geq 2\lambda_{\max}(\tilde{B}_F)/(3\tilde{\kappa}), \end{aligned} \tag{62}$$

where the last inequality is obtained by (41) and the assumption that $\delta \leq 1/(8C_{\text{upper}}\kappa(B))$.

We then derive a lower bound of $||y(F)||_{\tilde{B}_F}$. By the triangle inequality, we have

$$\begin{aligned} ||y(F)||_{\tilde{B}_F} &\geq ||\tilde{v}||_{\tilde{B}_F} - ||\tilde{v} - y(F)||_{\tilde{B}_F} \geq \sqrt{\sum_{j=1}^{k'} \alpha_j^2} - \sqrt{\lambda_{\max}(\tilde{B}_F)||\tilde{v} - y(F)||_2} \\ &\geq \frac{1}{2}\sqrt{\sum_{j=1}^{k'} \alpha_j^2} + \frac{1}{2}\sqrt{\frac{\lambda_{\max}(\tilde{B}_F)}{\tilde{\kappa}}} - \sqrt{2\lambda_{\max}(\tilde{B}_F)\delta} \geq \frac{1}{2}\alpha_1, \end{aligned}$$

where the last inequality is due to (62).

We first get the first term $y(F)^T C_F \tilde{v}$ in (59),

$$\begin{aligned} y(F)^T C_F \tilde{v} &= y(F)^T \tilde{v} + \frac{\eta}{\rho} y(F)^T (\tilde{A}_F - \rho \tilde{B}_F)\tilde{v} \\ &= 1 - \delta + \frac{\eta}{\rho}\{\tilde{\lambda}_1(F) - \rho\} y(F)^T \tilde{B}_F \tilde{v} \\ &= 1 - \delta + \frac{\eta}{\rho}\{\tilde{\lambda}_1(F) - \rho\}\alpha_1 \frac{\boldsymbol{\xi}_1^T \tilde{B}_F \boldsymbol{\xi}_1}{||\boldsymbol{\xi}_1||_2} \\ &= 1 - \delta + \eta\alpha_1 \frac{\{\tilde{\lambda}_1(F) - \rho\}}{\rho} ||y(F)||_{\tilde{B}_F} \\ &\geq 1 - \delta + \frac{1}{2}\eta\alpha_1^2 \frac{(1-\gamma)\lambda_{\max}(\tilde{B}_F)\delta}{\alpha_1^2\tilde{\kappa} + \gamma\lambda_{\max}(\tilde{B}_F)\delta} \\ &\geq 1 - \delta + \frac{1}{2}\eta \frac{\alpha_1^2(1-\gamma)\delta}{\tilde{\kappa} + \gamma} \\ &\geq 1 - \delta + \frac{1}{3}\eta\lambda_{\max}(\tilde{B}_F)\frac{(1-\gamma)\delta}{\tilde{\kappa} + \gamma}, \end{aligned} \tag{63}$$

where the lower bound of $\frac{\tilde{\lambda}_1(F)-\rho}{\rho}$ in (63) is derived by

$$\begin{aligned} \frac{\tilde{\lambda}_1(F) - \rho}{\rho} &= \frac{\sum_{j=1}^{k'}\{\tilde{\lambda}_1(F) - \tilde{\lambda}_j(F)\}\alpha_j^2}{\sum_{j=1}^{k'}\tilde{\lambda}_j(F)\alpha_j^2} \geq \frac{\{\tilde{\lambda}_1(F) - \tilde{\lambda}_2(F)\}\sum_{j=2}^{k'}\alpha_j^2}{\tilde{\lambda}_1(F)\alpha_1^2 + \tilde{\lambda}_2(F)\sum_{j=2}^{k'}\alpha_j^2} \\ &= \frac{\{\tilde{\lambda}_1(F) - \tilde{\lambda}_2(F)\}||\tilde{v} - \alpha_1\boldsymbol{\xi}_1||_{\tilde{B}_F}^2}{\tilde{\lambda}_1(F)\alpha_1^2 + \tilde{\lambda}_2(F)||\tilde{v} - \alpha_1\boldsymbol{\xi}_1||_{\tilde{B}_F}^2} \geq \frac{(1-\gamma)\{\lambda_{\max}(\tilde{B}_F)/\tilde{\kappa}\}||\tilde{v} - \alpha_1\boldsymbol{\xi}_1||_2^2}{\alpha_1^2 + \gamma\{\lambda_{\max}(\tilde{B}_F)/\tilde{\kappa}\}||\tilde{v} - \alpha_1\boldsymbol{\xi}_1||_2^2} \\ &\geq \frac{(1-\gamma)\lambda_{\max}(\tilde{B}_F)\delta}{\alpha_1^2\tilde{\kappa} + \gamma\lambda_{\max}(\tilde{B}_F)\delta}, \end{aligned}$$

the last inequality is derived by the fact that

$$\delta \leq 2\delta - \delta^2 = (1-\delta)^2 + 1 - 2(1-\delta)y(F)^T\tilde{v} = ||\tilde{v} - (1-\delta)y(F)||_2^2 \leq ||\tilde{v} - \alpha_1\boldsymbol{\xi}_1||_2^2.$$

Then, we calculate the lower bound of the second term $||C_F\tilde{v}||_2^{-1}$ in (59).

$$\begin{aligned} ||C_F\tilde{v}||_2^2 &= ||\{I + \frac{\eta}{\rho}(\tilde{A}_F - \rho\tilde{B}_F)\}\tilde{v}||_2^2 = 1 + ||\frac{\eta}{\rho}\sum_{j=1}^{k'}\alpha_j\tilde{A}_F\boldsymbol{\xi}_j - \rho\sum_{j=1}^{k'}\alpha_j\tilde{B}_F\boldsymbol{\xi}_j||_2^2 \\ &= 1 + ||\sum_{j=1}^{k'}\alpha_j\frac{\eta}{\rho}\{\tilde{\lambda}_j(F) - \rho\}\tilde{B}_F\boldsymbol{\xi}_j||_2^2, \end{aligned} \tag{64}$$

followed by the fact $\tilde{v}^T\tilde{A}_F\tilde{v} - \rho\tilde{v}^T\tilde{B}_F\tilde{v} = 0$ and $\tilde{A}_F\boldsymbol{\xi}_j = \tilde{\lambda}_j(F)\tilde{B}_F\boldsymbol{\xi}_j$. Also, we can get the upper bound of the $\frac{\tilde{\lambda}_1(F)-\rho}{\rho}$,

$$\frac{\tilde{\lambda}_1(F) - \rho}{\rho} = \frac{\sum_{j=1}^{k'}\{\tilde{\lambda}_1(F) - \tilde{\lambda}_j(F)\}\alpha_j^2}{\sum_{j=1}^{k'}\tilde{\lambda}_j(F)\alpha_j^2} \leq \frac{\tilde{\lambda}_1(F)\sum_{j=1}^{k'}\alpha_j^2}{\tilde{\lambda}_1(F)\alpha_1^2} \leq \frac{2\lambda_{\max}(\tilde{B}_F)\delta}{\alpha_1^2} \leq 3\delta\tilde{\kappa}. \tag{65}$$

By the assumption that $\delta \leq \frac{1/\gamma - 1}{3C_{\text{upper}}\kappa(B)}$ and (65), we have

$$\tilde{\lambda}_2(F) \leq \rho \leq \tilde{\lambda}_2(F).$$

From the definition $||\tilde{v}||_2^2 = 1$, we can get $\alpha_1^2 \leq \lambda_{\max}(\tilde{B}_F)$. Thus, we have

$$\begin{aligned}
&||\sum_{j=1}^{k'} \alpha_j \frac{\eta}{\rho}\{\tilde{\lambda}_j(F) - \rho\}\tilde{B}_F\boldsymbol{\xi}_j||_2^2 \\
&\leq \alpha_1^2\{\tilde{\lambda}_j(F) - \rho\}^2\lambda_{\max}(\tilde{B}_F)(\frac{\eta}{\rho})^2 + \lambda_{\max}(\tilde{B}_F)\sum_{j=2}^{k'}\alpha_j^2(\frac{\eta}{\rho})^2\{\tilde{\lambda}_j(F) - \rho\}^2 \\
&\leq \lambda_{\max}^2(\tilde{B}_F)\eta^2(3\delta\tilde{\kappa})^2 + \lambda_{\max}(\tilde{B}_F)\eta^2\{\frac{\tilde{\lambda}_1(F)}{\rho} - 1\}^2\sum_{j=2}^{k'}\alpha_j^2 \\
&\leq \lambda_{\max}^2(\tilde{B}_F)\eta^2(3\delta\tilde{\kappa})^2 + 2\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta(3\delta\tilde{\kappa})^2 \\
&= 9\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2 + 18\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^3\tilde{\kappa}^2
\end{aligned} \tag{66}$$

Substituting results (66) to results (64), we have

$$\begin{aligned}
||C_F\tilde{v}||_2^2 &\leq 1 + 9\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2 + 18\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^3\tilde{\kappa}^2 \\
&\leq 1 + 12\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2.
\end{aligned} \tag{67}$$

The last inequality holds because the assumption $\delta \leq \frac{1}{8C_{\text{upper}}\kappa}$ and $\eta C_{\text{upper}}\lambda_{\max}(B) < 1$, thus, $2\delta \leq \frac{1}{4}$. Besides, from the fact that $\frac{1}{\sqrt{1+y}} \geq 1 - \frac{y}{2}$ for $|y| < 1$, we have

$$||C_F\tilde{v}||_2^{-1} \geq 1 - 6\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2 \tag{68}$$

Now, combining the result of $y(F)^T C_F\tilde{v}$ and $||C_F\tilde{v}||_2^{-1}$, we have

$$\begin{aligned}
y(F)^T v' &= y(F)^T C_F\tilde{v} \cdot ||C_F\tilde{v}||_2^{-1} \\
&\geq \{1 - \delta + \frac{1}{3}\eta\lambda_{\max}(\tilde{B}_F)\frac{(1-\gamma)\delta}{\tilde{\kappa}+\gamma}\}\{1 - 6\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2\} \\
&\geq 1 - \delta + \frac{1}{3}\eta\lambda_{\max}(\tilde{B}_F)\frac{(1-\gamma)\delta}{\tilde{\kappa}+\gamma} - 6\lambda_{\max}^2(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2 - 2\tilde{\kappa}^2\eta^3\lambda_{\max}(\tilde{B}_F)\delta^2\frac{(1-\gamma)\delta}{\tilde{\kappa}+\gamma} \\
&\geq 1 - \delta + \frac{1}{3}\eta\lambda_{\min}(\tilde{B}_F)\frac{(1-\gamma)\delta}{\tilde{\kappa}+\gamma} - 6.25\lambda_{\max}(\tilde{B}_F)\eta^2\delta^2\tilde{\kappa}^2 \\
&\geq 1 - \delta + \frac{1}{8}\eta\lambda_{\min}(\tilde{B}_F)\frac{(1-\gamma)\delta}{\tilde{\kappa}+\gamma},
\end{aligned} \tag{69}$$

in which the last two inequality holds by the assumption that $\eta\lambda_{\max}(\tilde{B}_F) < 1$ and the condition that

$$\frac{(1-\gamma)}{\tilde{\kappa}+\gamma} \geq 30\eta\lambda_{\max}(\tilde{B})\delta\tilde{\kappa}^2,$$

which implied by the following inequality based on the assumption 1:

$$\delta \leq \frac{1-\gamma}{30(1+c)C_{\text{upper}}^2\eta\lambda_{\max}(B)\kappa^2(C_{\text{upper}}\kappa+\gamma)}.$$

Finally, we get the lower bound of $y(F)^T v'$,

$$\langle y(F), v'\rangle \geq 1 - \delta + \frac{1+c}{8}\eta\lambda_{\min}(B)\{1 - y(F)^T\tilde{v}\}\frac{1-\gamma}{C_{\text{upper}}\kappa+\gamma}.$$

□

7 Proof of Theorem 4

Theorem 18 (Fell version of Theorem 4)**.** Under Theorem 3 and choose the stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B)+\gamma}} < \frac{1}{2},$$

and we further denote

$$\lambda_{gap} = \min_{j>1} \frac{\lambda_1 - (1+a)\lambda_j}{\sqrt{1+\lambda_1^2}\sqrt{1+(1-a)^2\lambda_j^2}} \tag{70}$$

as the eigengap for the GEP. Assume n is sufficiently large such that,

$$\begin{aligned} n \geq \tilde{\Omega}(\max\{&\frac{d}{\lambda_{gap}^2 \text{cr}^2(A,B)}, \frac{d}{(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(A,B))^2}, \frac{\sqrt{d\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \\ &\frac{\sqrt{d\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}, \frac{\sqrt{d\log\frac{1}{\zeta}}}{\text{cr}(A,B)\lambda_{gap}\sqrt{\rho}}, \frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{d\log\frac{1}{\zeta}}}{\theta(A,B)\sqrt{\rho}}\}) \end{aligned}$$

for given failure probability $\zeta > 0$. Then in Algorithm 3, if we set $m = O(\log n)$ and if the input vector v_0 with $\|v_0\|_2 = 1$ satisfying $\frac{|\langle v^*, v_0\rangle|}{\|v^*\|_2} \geq 1 - \frac{\theta(A,B)}{2}$ with $\theta(A,B)$ in (10). We have the following with probability at least $1-\zeta$,

$$1 - \frac{\langle v^*, v_m\rangle}{\|v^*\|_2} \leq O(\frac{\theta(A,B)}{(1-\nu)^2} \times (\frac{\rho^2(E_A)+\rho^2(E_B)}{\lambda_{gap}^2\text{cr}^2(A,B)} + \frac{1}{\lambda_{gap}^2\text{cr}^2(A,B)}\frac{d\log n\log d\log\frac{1}{\zeta}}{n^2\rho})).$$

Similarly, for the deterministic setting where $E_A = E_B = 0$ and $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and we set some appropriate parameters in Algorithm 3, with probability at least $1-\zeta$

$$1 - \frac{\langle \tilde{v}, v_m\rangle}{\|\tilde{v}\|_2} \leq \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2\text{cr}^2(A,B)}\frac{d\log\frac{1}{\zeta}}{n^2\rho}). \tag{71}$$

Proof. We mainly focus on the stochastic setting. The result for the deterministic setting is just a special case where $\tilde{\iota}(k') = 0$ since $E_A = E_B = 0$.

In each iteration t we denote the vector V^t as the solution of the following GEP.

$$V^t = \arg\max_{v\in\mathbb{R}^d} v^T\tilde{A}^t v, \text{ s.t. } v^T\tilde{B}^t v = 1. \tag{72}$$

We also denote $y^* = \frac{v^*}{\|v^*\|_2}$ and $y^t = \frac{V^t}{\|V^t\|_2}$. The following theorem shows the error of the parameter for each iteration.

Theorem 19. Under Theorem 3 and choose the stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B)+\gamma}} < \frac{1}{2}.$$

Suppose in the t-th iteration, v_{t-1} satisfies $\langle y^t, v_{t-1}\rangle \geq 1 - \theta(A,B)$, where

with γ in Theorem 3. Then we have

$$\sqrt{1-|\langle y^*, v_t\rangle|} \leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + 2\sqrt{1-|\langle y^t, y^*\rangle|}. \tag{73}$$

To proof Theorem 4 we will show the following three lemmas:

Lemma 14. Under Theorem 19, if $\rho(E_A) + \rho(E_B) \leq O(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(A,B))$ and $n \geq \Omega(\frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{\min\{\theta(A,B),\sqrt{\theta(A,B)}\}\sqrt{\rho}})$, then for all $t\in[m]$, if $|\langle v^*, v_{t-1}\rangle|/\|v^*\|_2 \geq 1 - \frac{\theta(A,B)}{2}$, then

$$\langle y^t, v_{t-1}\rangle \geq 1 - \theta(A,B). \tag{74}$$

Proof of Lemma 14. Since

$$\langle y^t, v_{t-1}\rangle \geq \frac{|\langle v^*, v_{t-1}\rangle|}{\|v^*\|_2} - \|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2,$$

it is sufficiently to show that $\|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 \leq \frac{\theta(A,B)}{2}$.

To show this, we will provide a stronger statement.

Lemma 15. Under Theorem 3, for each $t \in [m]$ if $n \geq \Omega(\frac{\sqrt{d\log d\log\frac{m}{\zeta}}}{\mathrm{cr}(A,B)\sqrt{\rho}})$,

$$\|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 \leq O(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\mathrm{cr}(A,B)} + \frac{1}{\lambda_{gap}\mathrm{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}).$$

Proof. By Theorem 4.3 in [Stewart, 1979] we have

$$\frac{\|v^t - v^*\|_2}{\|v^*\|_2} \leq \frac{\iota}{\Delta\tilde{\lambda}^t\mathrm{cr}(\tilde{A}^t, \tilde{B}^t)},$$

where $\iota = \sqrt{\rho(E_A + Z_1^t)^2 + \rho(E_B + Z_2^t)^2}$ and $\Delta\tilde{\lambda}^t = \min_{k>1}\chi\{\lambda_1, \tilde{\lambda}_k^t\} > 0$ with

$$\chi\{\lambda_1, \tilde{\lambda}_k^t\} = \frac{|\lambda_1 - \tilde{\lambda}_k^t|}{\sqrt{1+\lambda_1^2}\sqrt{1+\tilde{\lambda}_k^{t2}}}.$$

Thus, we have

$$\begin{aligned}\|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 &= \frac{1}{\|v^t\|_2\|v^*\|_2}\|v^t\|v^*\|_2 - v^*\|v\|_2\|_2 \\ &\leq \frac{2}{\|v^*\|_2}\|v^t - v^*\|_2 \\ &\leq \frac{2}{\|v^*\|_2}\|v^t - v^*\|_2 \leq 2\frac{\iota}{\Delta\tilde{\lambda}^t\mathrm{cr}(\tilde{A}^t, \tilde{B}^t)}.\end{aligned}$$

By Theorem 3 we can see that $\Delta\tilde{\lambda} \geq \lambda_{gap}$, and by Theorem 2.4 in [Stewart, 1979] we have $\mathrm{cr}(\tilde{A}^t, \tilde{B}^t) \geq \mathrm{cr}(A,B) - \iota$. Thus we have

$$\|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 \leq O(\frac{\iota}{\lambda_{gap}(\mathrm{cr}(A,B) - \iota)}).$$

By the definition of ι we have $\iota \leq \rho(E_A) + \rho(E_B) + O(\frac{\sqrt{dT\log d\log\frac{1}{\zeta}}}{n\sqrt{\rho}})$. Since $\frac{\rho(E_A)+\rho(E_B)}{\mathrm{cr}(k')} \leq O(1)$ and $n \geq \Omega(\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{\mathrm{cr}(k')\sqrt{\rho}})$, we have

$$\|\frac{v^t}{\|v^t\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 \leq O(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\mathrm{cr}(A,B)} + \frac{1}{\lambda_{gap}\mathrm{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}).$$

□

Thus, we can see that when $\rho(E_A) + \rho(E_B) \leq O(\theta(A,B)\lambda_{gap}\mathrm{cr}(A,B))$ and $n \geq \Omega(\frac{1}{\lambda_{gap}\mathrm{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{\theta(A,B)\sqrt{\rho}})$ then $\|\frac{v^t(F)}{\|v^t(F)\|_2} - \frac{v^*}{\|v^*\|_2}\|_2 \leq \frac{\theta(A,B)}{2}$. □

Lemma 16. Under Theorem 3 and Lemma 16, for all $t \in [m]$, if $\langle y^t, v_{t-1}\rangle \geq 1 - \theta(A,B)$, then

$$\sqrt{1 - |\langle y^*, v_t\rangle|} \leq \nu\sqrt{1 - |\langle y^*, v_{t-1}\rangle|} + O(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\mathrm{cr}(A,B)} + \frac{1}{\lambda_{gap}\mathrm{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \tag{75}$$

and

$$\frac{|\langle v^*, v_t\rangle|}{\|v^*\|_2} \geq 1 - \frac{\theta(A,B)}{2}. \tag{76}$$

Combining these two lemmas we have under the assumptions, for each $t \in [m]$,

$$\sqrt{1-\frac{|\langle v^*, v_m\rangle|}{\|v^*\|_2}} \leq \nu^m \sqrt{\theta(A,B)} + O(\frac{1}{1-\nu}(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\text{cr}(A,B)} + \frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}})). \tag{77}$$

Take $m = O(\log n)$ we have the proof.

Proof of Lemma 16 . By Theorem 19 and Lemma 11 we have

$$\begin{aligned}
\sqrt{1-|\langle y^*, v_t\rangle|} &\leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + 2\sqrt{1-|\langle y^t, y^*\rangle|} \\
&\leq \nu\sqrt{1-|\langle y^*, v_{t-1}\rangle|} + O(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\text{cr}(A,B)} + \frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \\
&\leq \nu\sqrt{\theta(A,B)} + O(\frac{\rho(E_A)+\rho(E_B)}{\lambda_{gap}\text{cr}(A,B)} + \frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{n\sqrt{\rho}}) \\
&\leq \frac{\sqrt{\theta(A,B)}}{\sqrt{2}},
\end{aligned}$$

where the last inequality is due to that $\nu < \frac{1}{2}$, $\rho(E_A)+\rho(E_B) \leq O(\max\{\theta(A,B), \sqrt{\theta(A,B)}\}\lambda_{gap}\text{cr}(A,B))$ and $n \geq \Omega(\frac{1}{\lambda_{gap}\text{cr}(A,B)}\frac{\sqrt{dm\log d\log\frac{m}{\zeta}}}{\min\{\theta(A,B),\sqrt{\theta(A,B)}\}\sqrt{\rho}})$. □

□

7.1 Proof of Theorem 19

The proof of Theorem 3 is almost the same as the proof of Theorem 13, we omit it here.

The same as Lemma 12, we have the following lemma. In each iteration t we denote the vector v^t as the solution of the following GEP.

$$v^t = \arg\max_{v\in\mathbb{R}^d} v^T\tilde{A}^t v, \text{ s.t. } v^T\tilde{B}^t v = 1. \tag{78}$$

Lemma 17. Assume that Theorem 3 holds, consider in the t-th iteration, let $y^t = \frac{v^t}{||v^t||_2}$, and $y^* = \frac{v^*}{||v^*||_2}$. Given any $\tilde{v}$, such that $||\tilde{v}||_2 = 1$ and $\tilde{v}^T y > 0$. Let $\rho = \frac{\tilde{v}^T\tilde{A}^t\tilde{v}}{\tilde{v}^T\tilde{B}^t\tilde{v}}$, and let $v' = \frac{C^t\tilde{v}}{||C^t\tilde{v}||_2}$, where

$$C^t = I + \frac{\eta}{\rho}(\tilde{A}^t - \rho\tilde{B}^t),$$

and $\eta > 0$ is a positive constant. Define η such that

$$\eta\lambda_{\max}(B) < \frac{1}{1+c}.$$

Denote $\delta = 1 - \langle y^t, \tilde{v}\rangle$, and $1-\delta \geq 1-\theta(A,B)$, where

$$\theta(A,B) = \min\{\frac{1}{8C_{\text{upper}}\kappa(B)}, \frac{1/\gamma-1}{3C_{upper}\kappa(B)}, \frac{1-\gamma}{30(1+c)C_{upper}^2\eta\lambda_{\max}(B)\kappa^2(B)\{C_{\textbf{upper}}\kappa(B)+\gamma\}}\}.$$

Then, under the conditions in Theorem 13 we have

$$v'^T y^t \geq \tilde{v}^T y^t + \frac{1+c}{8}\eta\lambda_{\min}(B)\{1-\tilde{v}^T y^t\}\frac{1-\gamma}{C_{\text{upper}}\kappa+\gamma}. \tag{79}$$

The proof is the same as Lemma 12, where $F = \{1, 2, \cdots, d\}$.

Proof of Theorem 19. By Lemma 17, take $\tilde{v} = v_{t-1}$ with $v' = v_t$. Since by the assumption of in the t-th iteration, v_{t-1} satisfies $\langle y^t, v_{t-1}\rangle \geq 1 - \theta(A,B)$ we have

$$\langle v_t, y^t\rangle \geq \langle v_{t-1}, y^t\rangle + \frac{1+c}{8}\eta\lambda_{\min}(B)\{1 - \langle v_{t-1}, y^t\rangle\}\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}. \tag{80}$$

Thus, we have

$$1 - \langle v_t, y^t\rangle \leq (1 - \langle v_{t-1}, y^t\rangle)\{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}\}. \tag{81}$$

That is

$$\|v_t - y^t\|_2 \leq \|v_{t-1} - y^t\|_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}}.$$

$$\begin{aligned}
||y^* - v_t||_2 &\leq ||y^t - v_t||_2 + ||y^t - y^*||_2 \\
&\leq ||y^t - v_{t-1}||_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + ||y^t - y^*||_2 \\
&\leq ||y^* - v_{t-1}||_2\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + 2||y^t(F) - y^*||_2.
\end{aligned}$$

Thus, we have

$$\sqrt{1 - |\langle y^*, v_t\rangle|} \leq \sqrt{1 - |\langle y^*, v_{t-1}\rangle|}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa + \gamma}} + 2\sqrt{1 - |\langle y^t, y^*\rangle|}.$$

□

8 Proof of Theorem 5

Theorem 20 (Full version of Theorem 5)**.** The optimization problem (16) is (ϵ, δ)-DP under Assumption 1 for $0 < \epsilon, \delta < 1$. Moreover, under Assumption 2 and assume that $\|E_A\|_{\infty,\infty}, \|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, if n is sufficiently large such that $n \geq \Omega(\max\{\frac{\lambda_{\max}^2(B)\lambda_1^2 d^3 \log\frac{1}{\zeta}\log d}{\rho\theta(A,B)\Delta_{gap}^2}, \frac{\sqrt{d\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$. Then set $\phi = \tilde{O}(\lambda_{\max}(B)\lambda_1(\frac{\sqrt{d}}{\sqrt{n}} + \frac{\sqrt{d}}{n\sqrt{\rho}}))$ and $K = 1$ in (16) we have with probability at least $1 - \zeta$

$$\langle v_0, v^*\rangle \geq 1 - \theta(A,B)/2,$$

where for a matrix A, $\|A\|_{\infty,\infty}$ is defined as the maximal absolute value of all the entries in A, v_0 is the leading eigenvector of $\hat{P}$ and v^* is the leading general eigenvector of (A,B).

We proof the following theorem instead. Note that the proof of Theorem 20 is just based on Theorem 21 with the bound $\|\tilde{A} - \bar{A}\|_{\infty,\infty} = \tilde{O}(\lambda_{\max}(B)\lambda_1(\frac{\sqrt{d}}{\sqrt{n}} + \frac{\sqrt{d}}{n\sqrt{\rho}}))$.

Theorem 21. The optimization problem (16) is (ϵ, δ)-DP under Assumption 1 for $0 < \epsilon, \delta < 1$. Moreover, under Assumption 2 and assume that $\|E_A\|_{\infty,\infty}, \|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, if n is sufficiently large such that $n \geq \Omega(\max\{\frac{\lambda_{\max}^2(B)\lambda_1^2 d^3 \log\frac{1}{\zeta}\log d}{\rho\theta(A,B)\Delta_{gap}^2}, \frac{\sqrt{d\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$. Then set $\phi > 2\|\tilde{A} - \bar{A}\|_{\infty,\infty}$ and $K = 1$ in (16) we have with probability at least $1 - \zeta$

$$\langle v_0, v^*\rangle \geq 1 - \theta(A,B)/2,$$

where for a matrix A, $\|A\|_{\infty,\infty}$ is defined as the maximal absolute value of all the entries in A, v_0 is the leading eigenvector of $\hat{P}$ and v^* is the leading general eigenvector of (A,B).

Proof of Theorem 21. The proof of (ϵ, δ)-DP is just followed by the Gaussian mechanism and Assumption 1. Next we focus on the utility.

Before that we define some additional notations. Let $V^* \in \mathbb{R}^{d\times d}$ be d generalized eigenvectors and let Λ^* be a diagonal matrix of generalized eigenvalues of the matrix pair (A, B). The matrix A can be rewritten in terms of its generalized eigenvectors and generalized eigenvalues up to sign jointly $A = BV^*\Lambda^*(V^*)^T B$ [Gao et al., 2015]. Let $\bar{A} = \tilde{B}V^*\Lambda^*(V^*)^T\tilde{B}$ and $P = V^*_{:,K}(V^*_{:,K})^T$, where $V^*_{:,K}$ are the first K generalized eigenvectors of (A, B). Then we have the the following theorem whose proof follows the proof of the Proposition 1 in [Tan et al., 2018]. We omit it here for simplicity.

Theorem 22. Assume n is sufficiently large such as $\rho(E_B) \le c\lambda_{\min}(B)$, where c is the same constant in Theorem 3. Let $\Delta_{gap} = \lambda_K - c\kappa(B)\lambda_{K+1}/(1-c) > 0$. Then we have the following if $\phi > 2\|\tilde{A} - \bar{A}\|_{\infty,\infty}$, then the solution $\hat{P}$ in (16) satisfies

$$\|\hat{P} - P^*\|_F \le C(\frac{d}{\Delta_{gap}}\|\tilde{A} - \bar{A}\|_{\infty,\infty} + K\|\tilde{B} - B\|_2), \tag{82}$$

where $C > 0$ is a constant.

We will use Theorem 22 to proof our main result.

For $\|\tilde{B} - B\|_2$ by the definition we have the following with probability at least $1 - \zeta$

$$\|\tilde{B} - B\|_2 \le \|Z_2\|_2 + \rho(E_B) \le O(\frac{\sqrt{d\log\frac{1}{\zeta}}}{n\sqrt{\rho}} + \rho(E_B)).$$

For $\|\tilde{A} - \bar{A}\|_{\infty,\infty}$ we have with probability as least $1 - 3\zeta$

$$\begin{aligned}
&\|\tilde{A} - \bar{A}\|_{\infty,\infty} \le \|Z_1\|_{\infty,\infty} + \|\hat{A} - \bar{A}\|_{\infty,\infty} \\
&\le \|Z_1\|_{\infty,\infty} + \|\hat{A} - A\|_{\infty,\infty} + \|A - \bar{A}\|_{\infty,\infty} \\
&= \|Z_1\|_{\infty,\infty} + \|E_A\|_{\infty,\infty} + \|BV^*\Lambda^*(V^*)^T B - \tilde{B}V^*\Lambda^*(V^*)^T B\|_{\infty,\infty} \\
&+ \|\tilde{B}V^*\Lambda^*(V^*)^T B - \tilde{B}V^*\Lambda^*(V^*)^T\tilde{B}\|_{\infty,\infty} \\
&\le \|Z_1\|_{\infty,\infty} + \|E_A\|_{\infty,\infty} + \|(Z_2 + E_B)V^*\Lambda^*(V^*)^T B\|_{\infty,\infty} \\
&+ \|\tilde{B}V^*\Lambda^*(V^*)^T(Z_2 + E_B)\|_{\infty,\infty} \\
&\le \|Z_1\|_{\infty,\infty} + \|E_A\|_{\infty,\infty} + (2+c)\sqrt{d}\lambda_{\max}(B)\lambda_1(\|Z_2\|_{\infty,\infty} + \|E_B\|_{\infty,\infty}) \\
&\le O(\|E_A\|_{\infty,\infty} + c\lambda_{\max}(B)\lambda_1\sqrt{d}\|E_B\|_{\infty,\infty} + c\lambda_{\max}(B)\lambda_1\frac{\sqrt{d\log d\log\frac{1}{\zeta}}}{n\sqrt{\rho}}),
\end{aligned}$$

where the third inequality is due to the following result, whose proof is almost the same as the proof of Theorem 3.

Theorem 23. Under Assumption 2, given any failure probability $\zeta > 0$, if n is sufficiently large such that, $n \ge \Omega(\frac{\sqrt{d\log d\log\frac{1}{\delta}\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\epsilon})$. Then with probability at least $1 - \zeta$,

$$(1-c)\lambda_j(B) \le \lambda_j(\tilde{B}) \le (1+c)\lambda_j(B) \tag{83}$$

where c is the constant as in Theorem 3.

Thus, in total we have with probability at least $1 - \zeta$

$$\begin{aligned}
\|\hat{P} - P^*\|_F \le O(\frac{d}{\Delta_{gap}}(&\|E_A\|_{\infty,\infty} + \sqrt{d}\lambda_{\max}(B)\lambda_1\|E_B\|_{\infty,\infty} \\
&+ \lambda_{\max}(B)\lambda_1\frac{\sqrt{d\log d\log\frac{1}{\zeta}}}{n\sqrt{\rho}}) + K(\frac{\sqrt{d\log\frac{1}{\zeta}}}{n\sqrt{\rho}} + \rho(E_B))).
\end{aligned} \tag{84}$$

Thus, take $K = 1$ we have

$$\langle v_0, v^*\rangle \ge 1 - O(\frac{d}{\Delta_{gap}}(\|E_A\|_{\infty,\infty} + \sqrt{d}\lambda_{\max}(B)\lambda_1\|E_B\|_{\infty,\infty} + \lambda_{\max}(B)\lambda_1\frac{\sqrt{d\log d\log\frac{1}{\zeta}}}{n\sqrt{\rho}}) + \rho(E_B))^2. \tag{85}$$

Thus, when $\|E_A\|_{\infty,\infty}, \|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$ and $\rho(E_B) = \sqrt{\frac{d}{n}}$ and $n \geq \Omega(\frac{\lambda_{\max}^2(B)\lambda_1^2 d^3 \log d \log 1/\zeta}{\rho\theta(A,B)\Delta_{gap}^2})$ then $\langle v_0, v^* \rangle \geq 1 - \theta(A,B)/2$ with probability at least $1-\zeta$. □

9 Proof of Theorem 9 and 10

In the following we will focus on the lower bound. Since our lower bound will be in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing its private version. More details can be found in [Barber and Duchi, 2014].

Let $\mathcal{P}$ be a class of distributions over a data universe $\mathcal{X}$. For each distribution $p \in \mathcal{P}$, there is a deterministic function $\theta(p) \in \Theta$, where Θ is the parameter space. Let $\phi : \Theta \times \Theta :\mapsto \mathbb{R}_+$ be a semi-metric function on the space Θ and $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (in this paper, we assume that $\phi(x,y) = \|x-y\|_2$ and $\Phi(x) = x^2$ unless specified otherwise). We further assume that $D = \{X_i\}_{i=1}^n$ are n i.i.d observations drawn according to some distribution $p \in \mathcal{P}$, and $\hat{\theta} : \mathcal{X}^n \mapsto \Theta$ be some estimator. Then the minimax risk in metric $\Phi \circ \phi$ is defined by the following saddle point problem:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \phi) := \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\Phi(\phi(\hat{\theta}(D), \theta(p))],$$

where the supremum is taken over distributions $p \in \mathcal{P}$ and the infimum over all estimators $\hat{\theta}$.

In the ϵ-DP or CDP model, the estimator $\hat{\theta}$ is obtained via some ϵ-DP or ρ-CDP mechanism Q. Thus, we can also define the ϵ-DP (ρ-CDP)-private minimax risk:

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) := \inf_{Q \in \mathcal{Q}} \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_{p,Q}[\Phi(\phi(\hat{\theta}(D), \theta(p))],$$

where $\mathcal{Q}$ is the set of all ϵ-DP mechanisms. Similarly, for ρ-CDP we can define

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \rho) := \inf_{Q \in \mathcal{Q}} \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_{p,Q}[\Phi(\phi(\hat{\theta}(D), \theta(p))],$$

Next, we recall two private Fano's Lemmas given in [Acharya et al., 2021, Kamath et al., 2021].

Lemma 18 (Theorem 2 in [Acharya et al., 2021])**.** Consider a set of distributions $\mathcal{V} = \{p_1, p_2, \cdots, p_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,

- $\Phi(\phi(\theta(p_i), \theta_(p_j)) \geq \alpha$,
- $D_{KL}(p_i, p_j) \leq \beta$, where D_{KL} is the KL-divergence,
- $D_{TV}(p_i, p_j) \leq \gamma$,

then we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) \geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \geq \max\{\frac{\alpha}{2}(1 - \frac{n\beta + \log 2}{\log M}), 0.4\alpha \min\{1, \frac{M}{e^{10\epsilon n\gamma}}\}\}. \tag{86}$$

Lemma 19. [Theorem 1.4 in [Kamath et al., 2021]] Consider a set of distributions $\mathcal{V} = \{p_1, p_2, \cdots, p_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,

- $\Phi(\phi(\theta(p_i), \theta_(p_j)) \geq \alpha$,
- $D_{KL}(p_i, p_j) \leq \beta$, where D_{KL} is the KL-divergence,
- $D_{TV}(p_i, p_j) \leq \gamma$,

then we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon, \delta) \geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \\ \geq \frac{\alpha}{2} \max\{1 - \frac{n\beta + \log 2}{\log M}, 1 - \frac{\rho(n^2\gamma^2 + n\gamma(1-\gamma)) + \log 2}{\log M}\}.$$

Thus, based on Lemma 18 and 19 it is sufficient for us to 1) find the appropriate metric $\Phi \circ \phi$. 2) Find a set of distributions $\{p_1, \cdots, p_M\}$. We first consider the metric. Here we use the subspace distance between two unit vectors in $\mathbb{R}^d$ as ρ and the squared function as $\Phi(\cdot)$.

Subspace distance For two unit vectors $v, w \in \mathbb{R}^d$, each of them defines a 1-dimensional subspace $\mathcal{S}$ and $\mathcal{S}'$. Then the distance between $\mathcal{S}$ and $\mathcal{S}'$ is defined as

$$\sin\Theta(\mathcal{S}, \mathcal{S}') = \|vv^T - ww^T\|_F,$$

where $\|\cdot\|_F$ is the Frobenious norm. For simplicity, we will overload notation and write $\sin\Theta(\mathcal{S}, \mathcal{S}') = \sin\Theta(v, w)$.

Proof of Theorem 9. For the packing set $\mathcal{V}$, we have the following lemma:

Lemma 20. [[Cai et al., 2013]] Let (Θ, ϕ) be a totally bounded metric space. For any subset $E \subset \Theta$, denote by $\mathcal{N}(E, \epsilon)$ the ϵ-covering number of E, that is, the minimal number of balls of radius ϵ whose union contained in E. Also denote by $\mathcal{M}(E, \epsilon)$ the ϵ-packing number of E, that is, the maximal number of points in E whose pairwise distance is at least ϵ. If there exist $0 \leq c_0 \leq c_1 < \infty$ and $d > 0$ such that:

$$(\frac{c_0}{\epsilon})^d \leq \mathcal{N}(\Theta, \epsilon) \leq (\frac{c_1}{\epsilon})^d$$

for all $0 < \epsilon \leq c_0$, then for any $1 \geq \alpha > 0$, there exists a packing set $\mathcal{V} = \{v_1, \cdots, v_m\}$ with $m \geq (\frac{c_0}{\alpha c_1})^d$ such that $\alpha\epsilon \leq \phi(v_i, v_j) \leq 2\epsilon$ for each $i \neq j$.

Now, for the set of all 1-dimensional subspaces $\mathbb{G}_{d,1}$ we have the following lemma regarding the metric entropy (due to [Szarek, 1982]).

Lemma 21. For any 1-dimensional subspaces $\mathcal{S}, \mathcal{S}' \in \mathbb{G}_{d,1}$, denoting its projection matrix vv^T and ww^T with $\|v\|_2 = \|w\|_2 = 1$, and define the metric on $\mathbb{G}_{d,1}$ by the subspace distance between $\mathcal{S}$ and $\mathcal{S}'$. Then for any $\epsilon \in (0, \sqrt{2})$,

$$(\frac{c_0}{\epsilon})^{d-1} \leq \mathcal{N}(\mathbb{G}_{d,1}, \epsilon) \leq (\frac{c_1}{\epsilon})^{d-1},$$

where c_0, c_1 are absolute constants.

By Lemmas 20 and 21, we know that there exists a packing set $\mathcal{V}$ with $\log|\mathcal{V}| \geq (d-1)\log\frac{c_0}{\alpha c_1}$ with $2\epsilon_1 \geq \phi(\mathcal{S}, \mathcal{S}') \geq \alpha\epsilon_1$, where $\mathcal{S}, \mathcal{S}' \in \mathbb{G}_{d,1}$ and α and ϵ_1 will be specified later. Since each subspace $\mathcal{S}$ corresponds to a projection matrix vv^T with $\|v\|_2 = 1$. Thus we can rewrite $\mathcal{V} = \{v_1, \cdots, v_m\}$.

Now we construct the collection of distributions; for each $v \in \mathcal{V}$, we define

$$\Sigma_v = \frac{\lambda}{5d(\lambda+1)} vv^T + \frac{1}{5d(\lambda+1)} I_d. \tag{87}$$

That is, $\lambda_1 = \frac{1}{5d}$ and $\lambda_2 = \cdots = \lambda_d = \frac{1}{5d(\lambda+1)}$. Then we let p_{v_i} denote the distribution $\mathcal{N}(0, \Sigma_{v_i})$.

Now, we first show that the distribution Σ_v satisfies Assumption 1 with high probability. For $x \sim \mathcal{N}(0, \Sigma_v)$, we know that there exists an orthogonal matrix $M \in \mathbb{R}^{d\times d}$ which satisfies $Mx \sim \mathcal{N}(0, \mathrm{Diag}(\Sigma_v))$, where

$$\mathrm{Diag}(\Sigma_V) = \begin{bmatrix} \frac{1}{5d} & & & \\ & \frac{1}{5d(\lambda+1)} & & \\ & & \ddots & \\ & & & \frac{1}{5d(\lambda+1)} \end{bmatrix}.$$

Thus, we have $\|x\|_2^2 = \|Mx\|_2^2 \sim \frac{1}{5d}\chi_k^2 + \frac{1}{5d(\lambda+1)}\chi_{d-1}^2$. For the χ^2-distribution, we have the following concentration bound:

Lemma 22 ([Laurent and Massart, 2000])**.** If $z \sim \chi_n^2$, then

$$\mathbb{P}[z - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x).$$

By Lemma 12, we have the following with probability at least $1 - \exp(-1) - \exp(-(p-1))$, $\|x\|_2^2 \leq \frac{1}{5d}5 + \frac{1}{5d(\lambda+1)}5(d-1) \leq 1$. Thus, $\|x\|_2 \leq 1$ with high probability.

The following lemma shows that the Total Variation distance between P_V and $P_{V'}$ can be bounded by the subspace distance between V and V'.

Lemma 23. For any pair of $v, v' \in \mathcal{V}$, by the KL-distance $D_{KL}(\cdot||\cdot)$ of two Gaussian distributions, we have that

$$D_{KL}(p_v||p_{v'}) \leq \frac{\lambda^2}{2(1+\lambda)}\|\sin\Theta(v, v')\|_F^2.$$

Thus, by Pinsker's inequality that is $\|p_v - p_{v'}\|_{TV}^2 \leq \frac{\lambda^2}{1+\lambda}\|\sin\Theta(v, v')\|_F^2$.

Proof of Lemma 23.

$$\begin{aligned} D(P_V||P_{V'}) &= D(\mathcal{N}(0, \Sigma_V)||\mathcal{N}(0, \Sigma_{V'})) \\ &= \frac{1}{2}\text{trace}(\Sigma_{V'}^{-1}(\Sigma_V - \Sigma_{V'})). \end{aligned}$$

Now

$$\Sigma_{V'}^{-1} = 5p(\lambda+1)[(1+\lambda)^{-1}V'V'^T + (I_p - V'V'^T)]$$

and

$$\Sigma_V - \Sigma_{V'} = \frac{\lambda}{5p(\lambda+1)}(VV^T - V'V'^T).$$

we can get

$$\text{trace}(\Sigma_{V'}^{-1}(\Sigma_V - \Sigma_{V'})) = \frac{\lambda^2}{1+\lambda}\|\sin\Theta(V, V')\|_F^2.$$

□

Thus, by Lemma 18 we have

$$\begin{aligned} \mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) &\geq \frac{1}{M}\sum_{i\in[M]} \mathbb{E}_{X\sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \\ &\geq \Omega(\max\{(\alpha\epsilon_1)^2(1 - \frac{n\frac{4\epsilon_1^2\lambda^2}{2(1+\lambda)} + \log 2}{(d-1)\log\frac{c_0}{\alpha\epsilon_1}}), (\alpha\epsilon_1)^2\min\{1, \frac{(\frac{c_0}{\alpha\epsilon_1})^{d-1}}{e^{20\epsilon n\epsilon_1\sqrt{\frac{\lambda^2}{2(1+\lambda)}}}}\}\}). \end{aligned} \tag{88}$$

Take $\epsilon_1 = \tilde{O}(\min\{\sqrt{\frac{1+\lambda}{\lambda^2}}\sqrt{\frac{d}{n}}, \sqrt{\frac{1+\lambda}{\lambda^2}}\frac{d}{n\epsilon}\}$ and $\alpha = O(1)$, we have

$$\begin{aligned} \mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) \geq \frac{1}{M}\sum_{i\in[M]} \mathbb{E}_{X\sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \geq \\ \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}). \end{aligned}$$

Thus, we can see that for any ϵ-DP Algorithm, there must exists a distribution $p \in \{p_1, \cdots, p_m\}$ such that

$$\mathbb{E}_{D\sim p^n, \mathcal{A}}[\|\frac{v_{priv}v_{priv}^T}{\|v_{priv}\|_2^2} - v^*(v^*)^T\|_F^2] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}),$$

where v^* is the leading eigenvector of the covariance matrix of p. Note that $\|\frac{v_{priv}v_{priv}^T}{\|v_{priv}\|_2^2} - v^*(v^*)^T\|_F^2 = 2 - 2\langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle^2$. Thus,

$$\begin{aligned}
&\mathbb{E}_{D\sim p^n,\mathcal{A}}[\|\frac{v_{priv}v_{priv}^T}{\|v_{priv}\|_2^2} - v^*(v^*)^T\|_F^2] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}) \\
\iff &\mathbb{E}_{D\sim p^n,\mathcal{A}}[2 - 2\langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle^2] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}) \\
\iff &\mathbb{E}_{D\sim p^n,\mathcal{A}}[\langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle^2] \leq 1 - \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}) \\
\iff &\mathbb{E}_{D\sim p^n,\mathcal{A}}[\langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle] \leq \sqrt{1 - \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2})} \leq 1 - \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}) \\
\iff &\mathbb{E}_{D\sim p^n,\mathcal{A}}[1 - \langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d^2}{n^2\epsilon^2}).
\end{aligned}$$

Take $\lambda = O(1)$ we complete the proof.

For ρ-CDP, by Lemma 19 we have

$$\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\phi, \epsilon) &\geq \frac{1}{M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n,Q}[\Phi(\phi(Q(X), \theta(p_i))] \\
&\geq \Omega\left((\alpha\epsilon_1)^2)\max\{1 - \frac{n\frac{4\epsilon_1^2\lambda^2}{2(1+\lambda)} + \log 2}{(d-1)\log\frac{c_0}{\alpha\epsilon_1}}, 1 - \frac{\rho(n^2\frac{4\epsilon_1^2\lambda^2}{2(1+\lambda)} + n2\epsilon_1\sqrt{\frac{\lambda^2}{1+\lambda}}(1 - 2\epsilon_1\sqrt{\frac{\lambda^2}{1+\lambda}})) + \log 2}{(d-1)\log\frac{c_0}{\alpha\epsilon_1}}\}\right).
\end{aligned}$$

Take $\epsilon_1 = \tilde{O}(\min\{\sqrt{\frac{1+\lambda}{\lambda^2}}\sqrt{\frac{d}{n}}, \sqrt{\frac{1+\lambda}{\lambda^2}}\frac{\sqrt{d}}{n\sqrt{\rho}}\}$ and $\alpha = O(1)$, we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\phi, \epsilon) \geq \frac{1}{M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n,Q}[\Phi(\phi(Q(X), \theta(p_i))] \geq\geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d}{n^2\rho}).$$

Thus, the same as the previous ϵ-DP case we have

$$\mathbb{E}_{D\sim p^n,\mathcal{A}}[1 - \langle\frac{v_{priv}}{\|v_{priv}\|_2}, v^*\rangle] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{d}{n} + \frac{1+\lambda}{\lambda^2}\frac{d}{n^2\rho}).$$

□

Proof of Theorem 10. The construction of the class of distributions follows the idea presented in [Vu et al., 2013b]. For self-completeness, we rephrase below some important lemmas. See [Vu et al., 2013b] for the proofs.

Similar to the proof of Theorem 9, we consider the same class of distribution as in (87). Thus, the key step is to find a packing set $\mathcal{V}$ in $\mathbb{G}_{d,1}$. We denote $\mathbb{S}_{m,1} = \{v : \|v\|_2 = 1, v \in \mathbb{R}^m\}$. Thus we can see there is a one-to-one map between $\mathbb{G}_{m,1}$ and $\mathbb{S}_{m,1}$. The next lemma provides a general method for constructing such local packing sets.

Lemma 24 (Local Stiefel Embedding)**.** Let the function $A_\alpha : \mathbb{S}_{d-1,1} \mapsto \mathbb{S}_{d,1}$ be defined in block form as

$$A_\alpha(v) = \begin{bmatrix}(1-\alpha^2)^{\frac{1}{2}} \\ \alpha v\end{bmatrix} \tag{89}$$

for $0 \leq \alpha \leq 1$. If $v_1, v_2 \in \mathbb{S}_{d-1,1}$, then

$$\alpha^2(1-\alpha^2)\|v_1 - v_2\|_2^2 \leq \|\sin\Theta(A_\alpha(v_1), A_\alpha(v_2))\|_F^2 \leq \alpha^2\|v_1 - v_2\|_2^2.$$

We then need the following lemma:

Lemma 25 (Hypercube construction [Massart, 2007])**.** Let $e \leq d-1$ and $s - 1 \in [0, d-2]$. There exists a subset $\{v_1, \cdots, v_M\} \subset \mathbb{V}_{d-1,1}$ satisfying the following properties:

1. $\|v_i\|_0 \leq s-1, \forall i \in [M]$,
2. $\|v_i - v_j\|_2^2 \geq \frac{1}{4}$,
3. $\log M \geq \max\{c(s-1)[1+\log(d/(s-1))], \log d\}$, where $c \geq \frac{1}{30}$ is an absolute constant.

Now we construct the set of distributions. Consider $\{v_1, \cdots, v_M\} \subset \mathbb{V}_{d-1,1}$ in Lemma 25, for each v_i we consider the distribution $p_j = \mathcal{N}(0, \Sigma_{A_\alpha(v_i)})$ where

$$\Sigma_{A_\alpha(v_i)} = \frac{\lambda}{5d(\lambda+1)} A_\alpha(v)(A_\alpha(v))^T + \frac{1}{5d(\lambda+1)} I_d.$$

We can see that the leading eigenvector $A_\alpha(v)$ is s-sparse and if $x \sim \mathcal{N}(0, \Sigma_{A_\alpha(v_i)})$ then $\|x\|_2 \leq 1$ with high probability by Lemma 22. Moreover, by Lemma 23 we have $D_{KL}(p_v||p_{v'}) \leq \frac{\lambda^2}{(1+\lambda)}\alpha^2$ and $D_{TV}(p_v, p_{v'}) \leq \sqrt{\frac{\lambda^2}{(1+\lambda)}}\alpha$. Thus, by Lemma 18 we have

$$\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) &\geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \\
&\geq \Omega(\alpha^2(1-\alpha^2) \max\{(1 - \frac{n\frac{\alpha^2\lambda^2}{1+\lambda} + \log 2}{c(s-1)\log\frac{d}{s}}), \min\{1, \frac{(\frac{d}{s-1})^{c(s-1)}}{e^{10\epsilon n\alpha\sqrt{\frac{\lambda^2}{(1+\lambda)}}}}\}\}).
\end{aligned}$$

Thus, take $\alpha = O(\min\{\sqrt{\frac{\lambda+1}{\lambda^2}}\sqrt{\frac{s\log d}{n}}, \sqrt{\frac{1+\lambda}{\lambda^2}}\frac{s\log d}{n\epsilon}\})$ and since $n \geq \Omega(\frac{s\log d}{\epsilon})$, we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) \geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{s\log d}{n} + \frac{1+\lambda}{\lambda^2}\frac{(s\log d)^2}{n^2\epsilon^2}).$$

Thus, the same as in the proof of Theorem 9 we have

$$\mathbb{E}_{D \sim p^n, \mathcal{A}}[1 - \langle \frac{v_{priv}}{\|v_{priv}\|_2}, v^* \rangle] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{s\log d}{n} + \frac{1+\lambda}{\lambda^2}\frac{(s\log d)^2}{n^2\epsilon^2}).$$

For ρ-CDP, by Lemma 19 we have

$$\begin{aligned}
\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) &\geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \\
&\geq \Omega\left(\alpha^2(1-\alpha^2)\max\{1 - \frac{n\frac{\alpha^2\lambda^2}{1+\lambda} + \log 2}{c(s-1)\log\frac{d}{s}}), 1 - \frac{\rho(n^2\frac{\alpha^2\lambda^2}{1+\lambda} + n\alpha\sqrt{\frac{\lambda^2}{1+\lambda}}(1-\alpha\sqrt{\frac{\lambda^2}{1+\lambda}})) + \log 2}{c(s-1)\log\frac{d}{s}}\}\right).
\end{aligned}$$

Take $\epsilon_1 = O(\min\{\sqrt{\frac{1+\lambda}{\lambda^2}}\sqrt{\frac{s\log d}{n}}, \sqrt{\frac{1+\lambda}{\lambda^2}}\frac{\sqrt{s\log d}}{n\sqrt{\rho}}\})$, we have

$$\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi \circ \phi, \epsilon) \geq \frac{1}{M} \sum_{i \in [M]} \mathbb{E}_{X \sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \geq\geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{s\log d}{n} + \frac{1+\lambda}{\lambda^2}\frac{s\log d}{n^2\rho}).$$

Thus

$$\mathbb{E}_{D \sim p^n, \mathcal{A}}[1 - \langle \frac{v_{priv}}{\|v_{priv}\|_2}, v^* \rangle] \geq \Omega(\frac{1+\lambda}{\lambda^2}\frac{s\log d}{n} + \frac{1+\lambda}{\lambda^2}\frac{s\log d}{n^2\rho}).$$

□

10 Lower Bounds for DP-SIR

In this section we study the lower bound of SIR in the DP model.

Proof of Theorem 11. For any unit vector $v \in \mathbb{R}^d$ we consider a random variable x following a mixture of Gaussian distribution P_v for x, i.e., $P_v = \alpha\mathcal{N}(\frac{1-\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v) + (1-\alpha)\mathcal{N}(-\frac{\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)$, where $M_v = \frac{\alpha(1-\alpha)}{25d^2}vv^T$ and $\alpha \in (0,1)$. Equivalently, we have $\mathbb{P}(Y=1) = \alpha$, $\mathbb{P}(Y=2) = 1-\alpha$, $x|Y=1 \sim \mathcal{N}(\frac{1-\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)$ and $x|Y=2 \sim \mathcal{N}(-\frac{\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)$.

First we will show the instance satisfies Assumption 1 and 2 with high probability. Note that by the previous definition, when $Y=1$, then we have $x - \frac{1-\alpha}{5d}v \sim \mathcal{N}(0, \frac{1}{5d}I_d - M)$, thus the concentration property of Gaussian, we can see $\|x - \frac{1-\alpha}{5d}v\|_2 \le 1$ with high probability. Thus $\|x\|_2 \le 2$ with high probability. The same to the case when $Y=2$. Based on Theorem 1 we can see it satisfies Assumption 1 and 2 with high probability.

We can also easily see that

1. $\mathbb{E}_{P_v}[x] = 0$;
2. $A = \mathrm{cov}[\mathbb{E}[x|Y]] = \sum_{i=1}^2 \mathbb{P}[Y=i]\mathbb{E}[x|Y=i]\mathbb{E}[x|Y=i]^T - (\sum_{i=1}^2 \mathbb{P}[Y=i])(\sum_{i=1}^2 \mathbb{P}[Y=i])^T = \frac{\alpha(1-\alpha)}{25d^2}vv^T = M_v$
3. $\Sigma_x = \mathrm{cov}[\mathbb{E}[x|Y]] + \frac{1}{25d^2}I_d - M_v = \frac{1}{25d^2}I_d$.

Thus, the generalized eigenvector will be v for P_v, i.e., $\theta(P_v) = v$.

Next, for any pair of unit vectors v, v' we now compute the KL-divergence between P_v and $P_{v'}$.

Lemma 26. For for any pair of unit vectors v, v' we have

$$D_{KL}(P_v\|P_{v'}) \le \frac{3\lambda^2}{1-\lambda^2}\|v - v'\|_2^2, \tag{90}$$

where $\lambda = \alpha(1-\alpha)$

Proof. By the convexity of KL divergence and since P_v and $P_{v'}$ are mixture distributions, we have

$$\begin{aligned}
&D_{KL}(P_v\|P_{v'}) \le \alpha D_{KL}(\mathcal{N}(\frac{1-\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)\|\mathcal{N}(\frac{1-\alpha}{5d}v'), \frac{1}{25d^2}I_d - M_v') \\
&+ (1-\alpha)D_{KL}(\mathcal{N}(-\frac{\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)\|\mathcal{N}(-\frac{\alpha}{5d}v', \frac{1}{25d^2}I_d - M_{v'})).
\end{aligned}$$

We consider the first term first. By the KL-divergence formula for two multivariate Gaussian we have

$$\begin{aligned}
&D_{KL}(\mathcal{N}(\frac{1-\alpha}{5d}v, \frac{1}{25d^2}I_d - M_v)\|\mathcal{N}(\frac{1-\alpha}{5d}v'), \frac{1}{25d^2}I_d - M_{v'}) \\
&= \frac{1}{2}\{\mathrm{Tr}((\frac{1}{25d^2}I_d - M_{v'})^{-1}(\frac{1}{25d^2}I_d - M_v)) - p \\
&+ \log(\frac{\det(\frac{1}{25d^2}I_d - M_{v'})}{\det(\frac{1}{25d^2}I_d - M_v)} + \frac{(1-\alpha)^2}{25d^2}(v-v')^T(\frac{1}{25d^2} - M_{v'})^{-1}(v-v')\},
\end{aligned}$$

For the first term we have

$$\begin{aligned}
&\mathrm{Tr}((\frac{1}{25d^2}I_d - M_{v'})^{-1}(\frac{1}{25d^2}I_d - M_v)) - p = \mathrm{Tr}((\frac{1}{25d^2}I_d - M_{v'})^{-1}(M_{v'} - M_v)) \\
&= \mathrm{Tr}(I_d - \lambda v'v'^T)^{-1}(\lambda v'v'^T - \lambda vv^T))
\end{aligned}$$

Recall that

$$\begin{aligned}
&(I_d - \lambda v'v'^T)^{-1} = (I_d - v'v'^T + (1-\lambda)v'v'^T)^{-1} \\
&= I_d - v'v'^T + \frac{1}{1-\lambda}v'v'^T = I_d + \frac{\lambda}{1-\lambda}v'v'^T.
\end{aligned}$$

Thus we have

$$\begin{aligned}&\text{Tr}((I_d-\lambda v'v'^T)^{-1}(\lambda v'v'^T-\lambda vv^T))=\text{Tr}\{(I_d+\frac{\lambda}{1-\lambda}v'v'^T)(\lambda v'v'^T-\lambda vv^T))\}\\&=\lambda\text{Tr}(\frac{1}{1-\lambda}v'v'^T-vv^T-\frac{\lambda}{1-\lambda}v'v'^Tvv^T)\\&=\lambda\text{Tr}\{\frac{\lambda}{1-\lambda}(I_d-v'v'^Tvv^T)\}\leq\frac{2\lambda^2}{1-\lambda}\|v-v'\|_F^2.\end{aligned}$$

For the term $\log(\frac{\det(\frac{1}{25d^2}I_d-M_{v'})}{\det(\frac{1}{25d^2}I_d-M_v)}$ by the form of M_v and $M_{v'}$ we can see there eigenvalues are the same. Thus we have it is zero. For the third term we have

$$\begin{aligned}&\frac{(1-\alpha)^2}{25d^2}(v-v')^T(\frac{1}{25d^2}-M_{v'})^{-1}(v-v')=(1-\alpha)^2(v-v')^T(I_d-\lambda v'v'^T)^{-1}(v-v')\\&\leq\frac{(1-\alpha)^2}{1-\lambda}\text{Tr}((v-v')^T(v-v'))\leq\frac{(1-\alpha)^2}{1-\lambda}\|v-v'\|_F^2.\end{aligned}$$

Thus in total we have

$$D_{KL}(\mathcal{N}(\frac{1-\alpha}{5d}v,\frac{1}{25d^2}I_d-M_v)\|\mathcal{N}(\frac{1-\alpha}{5d}v'),\frac{1}{25d^2}I_d-M_{v'})\leq\frac{1}{2}\frac{2\lambda^2+(1-\alpha)^2}{1-\lambda}\|v-v'\|_F^2.$$

Similarly we have

$$D_{KL}(\mathcal{N}(-\frac{\alpha}{5d}v,\frac{1}{25d^2}I_d-M_v)\|\mathcal{N}(-\frac{\alpha}{5d}v',\frac{1}{25d^2}I_d-M_{v'}))\leq\frac{1}{2}\frac{2\lambda^2+\alpha^2}{1-\lambda}\|v-v'\|_F^2.$$

Thus we have

$$\begin{aligned}&D_{KL}(P_v\|P_{v'})\leq\left(\alpha\frac{2\lambda^2+(1-\alpha)^2}{1-\lambda}+(1-\alpha)\frac{2\lambda^2+\alpha^2}{1-\lambda}\right)\frac{1}{2}\|v-v'\|_F^2\\&=\frac{2\lambda^2+\alpha(1-\alpha)}{1-\lambda}\frac{1}{2}\|v-v'\|_F^2\\&\leq\frac{3\lambda^2}{1-\lambda^2}\|v-v'\|_F^2.\end{aligned}$$

□

Now we back to our proof since by Lemma 20 and 21, we know that there exists a packing set $\mathcal{V}$ with $\log|\mathcal{V}|\geq(d-1)\log\frac{c_0}{\alpha c_1}$ with $2\epsilon_1\geq\|v-v'\|_2\geq\alpha\epsilon_1$ for any $v,v'\in\mathcal{V}$. Thus by Lemma 19 we have

$$\begin{aligned}&\mathcal{M}_n(\theta(\mathcal{P}),Q,\Phi\circ\phi,\epsilon)\geq\frac{1}{M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n,Q}[\Phi(\phi(Q(X),\theta(p_i))]\\&\geq\Omega((\alpha\epsilon_1)^2)\max\{1-\frac{n\frac{3\epsilon_1^2\lambda^2}{(1-\lambda^2)}+\log 2}{(d-1)\log\frac{c_0}{\alpha\epsilon_1}},1-\frac{\rho(n^2\frac{3\epsilon_1^2\lambda^2}{(1-\lambda^2)}+2n\epsilon_1\sqrt{\frac{3\lambda^2}{1-\lambda^2}}(1-2\epsilon_1\sqrt{\frac{3\lambda^2}{1-\lambda^2}}))+\log 2}{(d-1)\log\frac{c_0}{\alpha\epsilon_1}}\}.\end{aligned}$$

Take $\epsilon_1=O(\min\{\sqrt{\frac{1-\lambda^2}{\lambda^2}}\sqrt{\frac{d}{n}},\sqrt{\frac{1-\lambda^2}{\lambda^2}}\frac{\sqrt{d}}{n\sqrt{\rho}}\})$ and $\alpha=O(1)$ we have for any ρ-CDP Algorithm, there must exists a distribution p such that

$$\mathbb{E}_{D\sim p^n,\mathcal{A}}[\|\frac{v_{priv}}{\|v_{priv}\|_2^2}-v^*\|_2^2]\geq\Omega(\frac{1-\lambda^2}{\lambda^2}\frac{d}{n}+\frac{1-\lambda^2}{\lambda^2}\frac{d}{n^2\rho}).$$

Thus, the same as the previous proof for PCA we have

$$\mathbb{E}_{D\sim p^n,\mathcal{A}}[1-\langle\frac{v_{priv}}{\|v_{priv}\|_2},v^*\rangle]\geq\Omega(\frac{1-\lambda^2}{\lambda^2}\frac{d}{n}+\frac{1-\lambda^2}{\lambda^2}\frac{d}{n^2\rho}).$$

□

Proof of Theorem 12. We first construct the set of distributions. Consider $\{v_1, \cdots, v_M\} \subset \mathbb{V}_{d-1,1}$ in Lemma 25, for each v_i we consider the distribution $P_{v_i} = \alpha\mathcal{N}(\frac{1-\alpha}{5d}A_\gamma(v_i), \frac{1}{25d^2}I_d - M_{v_i}) + (1-\alpha)\mathcal{N}(-\frac{\alpha}{5d}A_\gamma(v_i), \frac{1}{25d^2}I_d - M_{v_i})$, where $M_{v_i} = \frac{\alpha(1-\alpha)}{25d^2}A_\gamma(v_i)A_\gamma(v_i)^T$ and $\alpha, \gamma \in (0,1)$, $A_\gamma(\cdot)$ is defined in Lemma 24. Equivalently, we have $\mathbb{P}(Y=1) = \alpha$, $\mathbb{P}(Y=2) = 1-\alpha$, $x|Y=1 \sim \mathcal{N}(\frac{1-\alpha}{5d}A_\gamma(v_i), \frac{1}{25d^2}I_d - M_{v_i})$ and $x|Y=2 \sim \mathcal{N}(-\frac{\alpha}{5d}A_\gamma(v_i), \frac{1}{25d^2}I_d - M_{v_i})$. Similar to the low dimensional case, we can see if $x \sim P_{v_i}$ then it satisfies $\|x\|_2 \leq 1$ with high probability thus it satisfies Assumption 1 and 2. We can also easily see that

1. $\mathbb{E}_{P_v}[x] = 0$;
2. $A = \mathrm{cov}[\mathbb{E}[x|Y]] = \sum_{i=1}^2 \mathbb{P}[Y=i]\mathbb{E}[x|Y=i]\mathbb{E}[x|Y=i]^T - (\sum_{i=1}^2 \mathbb{P}[Y=i])(\sum_{i=1}^2 \mathbb{P}[Y=i])^T = \frac{\alpha(1-\alpha)}{25d^2}A_\gamma(v)A_\gamma(v)^T = M_v$
3. $\Sigma_x = \mathrm{cov}[\mathbb{E}[x|Y]] + \frac{1}{25d^2}I_d - M_v = \frac{1}{25d^2}I_d$.

Thus, we can see $A_\gamma(v_i)$ is the leading eigenvector for the instance P_{v_i} which satisfies $\|A_\gamma(v_i)\|_2 = 1$ and $\|A_\gamma(v_i)\|_0 \leq s$. Moreover, by Lemma 26 we have for any $v_i, v_j \in \mathbb{V}_{d-1,1}$ we have

$$D_{kl}(P_{v_i}\|P_{v_j}) \leq \frac{3\lambda^2}{1-\lambda^2}\|A_\gamma(v_i) - A_\gamma(v_j)\|_2^2 \leq \frac{6\lambda^2}{(1-\lambda^2)}\gamma^2.$$

Thus we have $D_{TV}(P_{v_i}\|P_{v_j}) \leq \sqrt{\frac{6\lambda^2}{(1-\lambda^2)}}\gamma$. Thus by Lemma 19 and 24 we have

$$\begin{aligned}
&\mathcal{M}_n(\theta(\mathcal{P}), Q, \Phi\circ\phi, \epsilon) \geq \frac{1}{M}\sum_{i\in[M]}\mathbb{E}_{X\sim p_i^n, Q}[\Phi(\phi(Q(X), \theta(p_i))] \\
&\geq \Omega\left(\gamma^2(1-\gamma^2)\max\{1 - \frac{n\frac{\gamma^2\lambda^2}{1-\lambda^2} + \log 2}{c(s-1)\log\frac{d}{s}}), 1 - \frac{\rho(n^2\frac{\gamma^2\lambda^2}{1-\lambda^2} + n\gamma\sqrt{\frac{\lambda^2}{1-\lambda^2}}(1-\gamma\sqrt{\frac{\lambda^2}{1-\lambda^2}})) + \log 2}{c(s-1)\log\frac{d}{s}}\}\right).
\end{aligned}$$

Take $\gamma = O(\min\{\sqrt{\frac{1-\lambda^2}{\lambda^2}}\sqrt{\frac{s\log d}{n}}, \sqrt{\frac{1-\lambda^2}{\lambda^2}}\frac{\sqrt{s\log d}}{n\sqrt{\rho}}\}$ we have the result.

□